Cui *et al. Intell. Robot.* **2025**, *5*, 119-42
DOI: 10.20517/ir.2025.08

Intelligence & Robotics

Review

Open Access

Task cognition and planning for service robots

Yongcheng Cui[1], **Ying Zhang**[2], **Cui-Hua Zhang**[2], **Simon X. Yang**[3]

[1]School of Mathematics and Statistics, Shandong Normal University, Jinan 250014, Shandong, China.
[2]School of Electrical Engineering, and Hebei Key Laboratory of Intelligent Rehabilitation and Neuromodulation, Yanshan University, Qinhuangdao 066004, Hebei, China.
[3]Advanced Robotics and Intelligent Systems Laboratory, University of Guelph, Guelph ON N1G 2W1, Canada.

Correspondence to: Dr. Ying Zhang, School of Electrical Engineering, and Hebei Key Laboratory of Intelligent Rehabilitation and Neuromodulation, Yanshan University, No. 438 West Hebei Avenue, Qinhuangdao 066004, Hebei, China. E-mail: yzhang@ysu.edu.cn

How to cite this article: Cui, Y.; Zhang, Y.; Zhang, C. H.; Yang, S. X. Task cognition and planning for service robots. *Intell. Robot.* **2025**, *5*, 119-42. https://dx.doi.org/10.20517/ir.2025.08

Received: 21 Sep 2024 **First Decision:** 20 Dec 2024 **Revised:** 18 Jan 2025 **Accepted:** 20 Jan 2025 **Published:** 7 Feb 2025

Academic Editor: Jianjun Ni **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

With the rapid development of artificial intelligence and robotics, service robots are increasingly becoming a part of our daily lives to provide domestic services. For robots to complete such services intelligently and with high quality, the prerequisite is that they can recognize and plan tasks to discover task requirements and generate executable action sequences. In this context, this paper systematically reviews the latest research progress in task cognition and planning for domestic service robots, covering key technologies such as command text parsing, active task cognition (ATC), multimodal perception, and action sequence generation. Initially, the challenges traditional rule-based command parsing methods face are analyzed, and the enhancement of robots' understanding of complex instructions through deep learning methods is explored. Subsequently, the research trends in ATC are introduced, discussing the ability of robots to autonomously discover tasks by perceiving the surrounding environment through visual and semantic features. The discussion then moves to the current typical methods in task planning, comparing and analyzing four common approaches to highlight their advantages and disadvantages in this field. Finally, the paper summarizes the challenges of existing research and the future directions for development, providing references for further enhancing the task execution capabilities of domestic service robots in complex home environments.

Keywords: Service robot, task cognition, task planning, robot action sequence generation

1. INTRODUCTION

The primary intention and aspiration of robotics research is to enable robots to replace or even surpass human performance in work-related tasks. Nowadays, domestic service robots are expected to do more than just simple tasks such as chatting, tutoring, or sweeping; they are also anticipated to perform complex tasks such as warming meals and tidying desks, effectively taking over various household chores. Achieving this goal requires robots to have robust task cognition and effective task planning, as depicted in Figure 1. To be specific, task cognition helps robots understand the specific requirements of their tasks, which involves generating structured task descriptions based on user commands or real-time perception, including task types and key details. Task planning involves creating an effective sequence of actions to complete these tasks based on set goals and environmental conditions. Only with accurate task understanding and sufficient access to task-related information can robots effectively plan and execute these tasks.

The ultimate goal of research related to task cognition and planning is to better assist service robots in completing tasks with high quality. These are two continuous processes that service robots must undergo, as shown in Figure 1. To achieve task cognition comparable to human butlers, current methods require robots not only to understand task requirements based on user requests but also to autonomously infer household tasks through perception of the home environment, reaching a level of diversified discovery and reasoning of service tasks. In response to user-specified tasks, i.e., reactive task cognition (RTC), robots need intelligent parsing capabilities to interpret user instructions. Specifically, since user commands are based on linguistic habits, accurately extracting necessary task information (such as task type, items, and context) from varied expressions of instructions presents a challenge to the robot's intelligent passive command parsing capabilities. Additionally, the rich task demands in intelligent home life require robots to autonomously recognize task requirements by understanding the home environment, known as active task cognition (ATC). This capability allows robots to go beyond merely receiving tasks passively via user commands, enabling them to engage in diversified task cognition. However, the home environment presents challenges with its complex variety of items and unstructured information, leading to redundant information during ATC. Simplifying this information to extract valuable task-related details and uncovering the constraints among items, tasks, and the environment poses significant challenges to robots' capability to represent the service environment and reason about tasks.

In the aspect of task planning, once robots have completed task cognition and identified task requirements, simply relying on a few predefined simple actions is insufficient for the complex execution steps required in household service tasks. Therefore, it is necessary to develop specialized complex task planning methods tailored for domestic robots that can formulate task sequences adaptable to the complex and dynamic nature of home environments. In terms of logicality in task planning, robotic tasks in home environments involve more than just simple actions such as "picking up" and "putting down"; they require combinations of multiple sub-tasks, each composed of multi-step action logic that must be executed in a specific logical order. Any logical flaws in task planning can lead directly to the failure of task execution by robots. Human domestic life experience, which serves as valuable a priori common knowledge in domestic services, aids in guiding robots to plan logically coherent action sequences. Moreover, in terms of environmental adaptability in task planning, the constantly changing positions and states of items in the home, along with various unexpected situations, pose challenges to the smooth execution of tasks. This demands that robots conduct specific analyses for specific problems, understand and adapt flexibly to the current state of the environment, and plan reasonable actions based on real-time changes. For domestic robots, adapting to various home environments and being able to adjust plans in real-time based on actual situations is key to successful task execution.

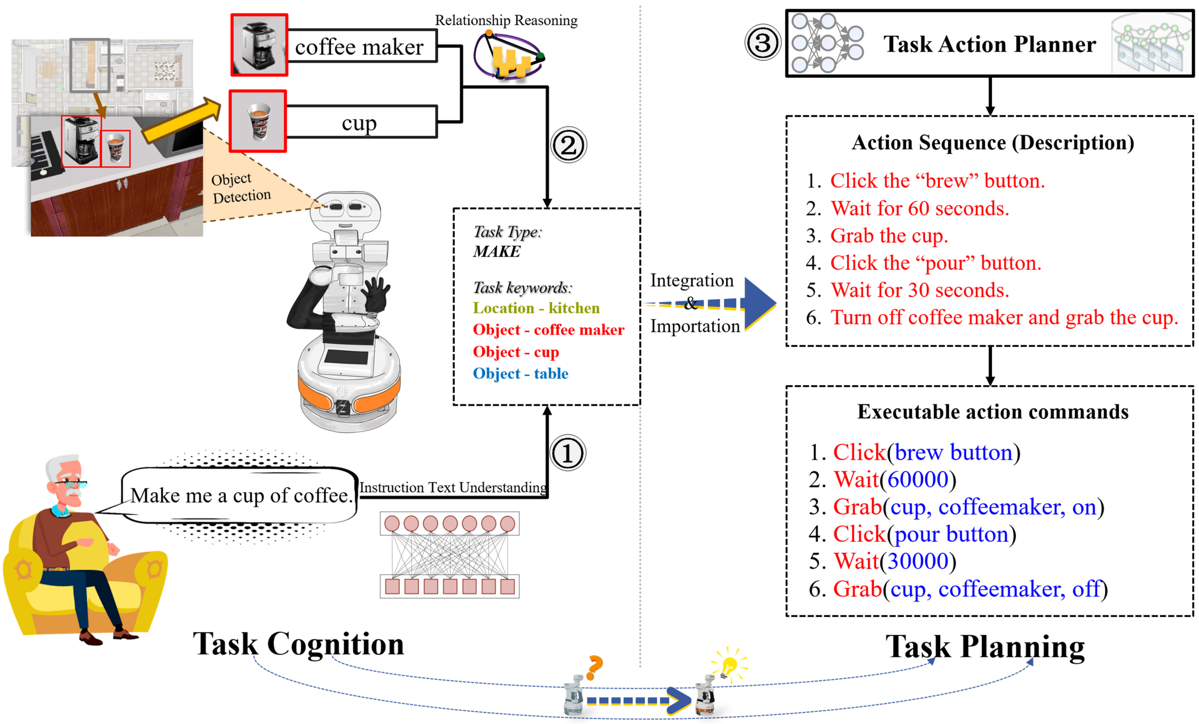


Figure 1. The relationship diagram between task cognition and task planning for service robots. Task cognition can be divided into two forms: RTC (①), where user text commands are parsed to extract task-related information, and ATC (②), where the environment is analyzed to infer potential tasks. Once the cognitive information is obtained, the task planner (③) then processes it to generate an executable action sequence. RTC: Reactive task cognition; ATC: active task cognition.

Existing reviews on robot cognition or planning often focus on industrial robots[1] and not specifically on service robots. Moreover, reviews related to intelligent robots typically concentrate on areas such as path planning[2,3] and scene understanding[4]. While Zhou *et al.* reviewed action planning algorithms for intelligent robots, they overlooked the latest trends in large model planning methods that have emerged prominently in the past two years[5]. Therefore, unlike traditional reviews on intelligent robots, this review specifically addresses task cognition and task planning methods for service robots, including current research progress, major challenges, and future development trends.

This paper focuses on task cognition and planning for home service robots, areas that are gaining increasing attention. While research has made progress, it often centers on individual technologies, resulting in a fragmented body of work. This review evaluates the feasibility and limitations of these approaches in real-world applications. As interest in this field grows, more research is expected to emerge, advancing its development.

The main contributions of this article are summarized as follows:
(1) A comprehensive analysis and review of the progress in robot task cognition.
(2) A survey of task planning methods in the robotics field.
(3) A discussion of the current issues in this area and possible future research directions.

2. METHODS OF TASK COGNITION IN ROBOTS

In current research, robot task cognition is categorized into two types: RTC and active. RTC, based on parsing user command texts, has already been extensively studied. However, research on ATC is not yet systematic. ATC relies on visual imagery to perceive and understand environmental information, granting robots a certain level of cognitive ability regarding the home environment, which is essential for implementing ATC. This section provides an overview of robot task cognition research, discussing both text-based and image-based cognition approaches.

2.1. Task cognition based on command texts

Currently, service robots must first understand user instructions to complete RTC[6]. The key process involves parsing command texts to identify the task type and extract key information. This is divided into two sub-tasks: recognizing task types and extracting essential details, with a focus on predicting user intent and extracting semantic concepts.

In task type recognition, the process essentially involves a text classification task aimed at assigning appropriate task labels to textual instructions. Due to the unstructured nature of text, extracting meaning presents certain challenges. Task labels can be assigned manually or automatically. As the scale of text data in applications grows, automatic text classification becomes increasingly important. Many researchers, both domestically and internationally, have made significant contributions to intent recognition, particularly in the field of natural language text classification.

In recent years, to overcome the time-consuming nature and limited applicability of manually designed features, neural network technology has been widely adopted. The core component of these methods is the use of machine learning-based embedding models, which map text into low-dimensional continuous feature vectors, eliminating the need for handcrafted features.

Specifically, word embedding models such as Word2Vec[7] and Glove[8] learn vector representations for words. These vectors are passed through feedforward layers, and classifiers such as logistic regression, Naive Bayes, or support vector machine (SVM)[9] classify task intent. The deep average network (DAN)[10] represents input text as the average of word vectors, offering simplicity, efficiency, and stability for short text classification, though it may lose word order information[11]. Recurrent neural networks (RNNs) capture word dependencies and text structure by treating text as a sequence. Kleenankandy *et al.* proposed an enhanced long short-term memory (LSTM) architecture, relational gated LSTM, which models relationships between sequence inputs[12]. They introduced the typed dependency tree-LSTM, utilizing sentence dependency structure to embed meaning, better capturing natural language syntax than traditional LSTMs. RNNs are trained to recognize temporal patterns, while convolutional neural networks (CNNs) focus on spatial patterns[13]. Ayetiran *et al.* proposed an attention-based CNN-embedded bidirectional LSTM model that captures bidirectional features and high-level semantics in opinion texts[14]. By combining CNN, which captures local patterns, with LSTM, which handles sequences, the model extracts richer features. Natural language instructions also include internal graph structures such as syntactic and semantic trees that define relationships between words.

With the rise of GPUs and greater computational power, transformer-based pre-trained language models (PLMs)[15] have gained popularity. These models are pre-trained on large unsupervised corpora and fine-tuned for downstream tasks. Compared to earlier CNN[16] and LSTM[17] models, Transformer-based PLMs have deeper structures and learn contextual representations by predicting words from large text corpora. Bidirectional encoder representations from transformers (BERT)[18] is one of the most widely used auto-encoding methods and is a PLM based on the transformer architecture. It captures contextual information through a bidirectional encoder, generating rich semantic vector representations, which are then classified for task intent. BERT's strength lies in its pre-training on massive text data, allowing it to learn robust language representations. This pre-training enables effective knowledge transfer for various natural language processing (NLP) tasks, significantly reducing the need for training from scratch on specific tasks. Text classification methods based on pre-trained models have greatly improved accuracy. Since the emergence of large language models (LLMs), represented by generative pre-trained transformer 4 (GPT-4)[19], there has been a surge in research and applications in the field of NLP due to their impressive artificial general intelligence-like (AGI-like) capabilities. LLMs can be divided into two categories based on their architecture: encoder-decoder (or encoder-only) and decoder-only structures. Encoder-decoder models (including BERT-style) generally offer stronger sequence learning and generation abilities compared to decoder-only models (including GPT-style), making them particularly suited for tasks such as machine translation, summarization, and chatbot systems[20]. However, encoder-only models have simpler structures, faster training and inference speeds, and are advantageous in tasks such as text classification or annotation.

In key information extraction, command parsing typically requires consideration of contextual and semantic information to extract the critical details embedded in the instructions, which differentiates it from general text classification tasks. Keyword extraction is essentially a slot-filling task, where the input word sequence is mapped to corresponding slot labels. Traditionally, intent recognition and slot filling have been handled separately, but this separation often leads to error propagation, causing cumulative mistakes. Therefore, the focus is on developing methods that jointly compute intent recognition and slot filling to improve overall parsing accuracy and robustness.

Previous work did not explicitly model the relationship between intent and slots, despite their strong interdependence. Liu *et al.* introduced a joint model combining an encoder-decoder with attention for intent prediction and slot filling[21]. Goo *et al.* proposed a method using LSTM and attention, with a gating mechanism to link intent and slot attention vectors, enhancing global optimization[22]. These models, however, rely heavily on large supervised datasets. Abro *et al.* addressed this by introducing the WFST-BERT model, which reduces the need for large data using a weighted finite-state transducer and BERT[23]. Qin *et al.* proposed a collaborative transformer model with co-interaction attention to model the interaction between intent detection and slot filling[24]. Some researchers introduced SlotRefine, a non-autoregressive model with a two-pass iterative mechanism to resolve slot inconsistencies[25]. To leverage the relationship between objects and commands in home service environments and address semantic gaps, He *et al.* proposed a multi-task learning intent detection system that combines a knowledge base with slot filling, enhancing performance through a weighted loss function[26]. Rajapaksha *et al.* developed a semantic parser integrated with a semantic-based ontology to handle unknown terms in high-level user commands[27]. This approach uses semantic web technology, enabling the robot to understand unknown terms by communicating with the parser, which represents additional concepts related to the target object.

Instruction parsing models, which are always based on slot filling, use attention mechanisms to assign weights to words, capturing their importance and generating sentence representations to improve classification performance. The input is the model's output hidden states (word vectors), with the attention layer output c^I as given in

$$c^I = \sum_{i=1}^{N} \alpha_i^I H_i \tag{1}$$

where N is the number of instruction words, H_i is the output hidden state of each BiGRU, and α_i^I is the attention weight for each word, as given in

$$\alpha_i^I = \frac{\exp(\sigma(W_{he}^I H_i))}{\sum_{k=1}^{N} \exp(\sigma(W_{he}^I H_k))} \tag{2}$$

where σ is the activation function, and W is the weight matrix. The attention layer enhances classification accuracy by focusing on key sentence parts, minimizing the impact of irrelevant or misleading information. The hidden states are combined with the attention output, and task classification is performed via a softmax layer to determine the instruction task type, as given in

$$y^I = soft\max(w_{hy}^I (H_N + c^I)) \tag{3}$$

where y^I is the task classification result, ${w_{hy}}^I$ is the weight matrix, and H_N is the model's final hidden state.

In summary, service task command parsing can be framed as a combination of text classification and sequence labeling, as shown in Figure 2. Rule-based approaches are suitable only for small data sets and fixed environments; as command complexity increases, the demand for manually defined rules grows significantly. In contrast, deep learning methods can effectively handle semantic information in context, making them more suitable for the dynamic nature of home environments. A major challenge is enhancing the generalizability of text parsing models due to diverse user expressions. Additionally, various object features can lead to matching issues between attributes and object names, making semantic relationships among extracted keywords and intent recognition important areas for research. Therefore, when robots receive user commands, employing word embedding mechanisms to capture command features and designing a specialized task classification and keyword extraction model for intelligent parsing are essential for accurately inferring user task requirements, which is crucial for successful task execution.

2.2. Task cognition based on visual images

Currently, cognition based on visual images is regarded as a highly challenging research area within pattern recognition and computer vision (CV). While most visual cognition focuses on understanding the environment and scenes, differing from the goal of ATC, its robust semantic understanding of the environment is crucial for implementing ATC. Both research areas require recognizing and processing information from images to infer deeper semantic content, achieving a level of understanding that surpasses the surface information presented in the images. Essentially, task cognition is a form of visual reasoning, where reasoning itself forms the foundation of the cognitive process, as illustrated in Figure 3.

Visual reasoning requires recognizing elements in images and understanding their interactions, meanings, and functions in the real world. Advancements in deep learning-driven object detection[28-33] have allowed researchers to achieve precise object categories and locations. Notably, faster region-based CNN (R-CNN)[28] and YOLO v9[30] exemplify the leading methods for efficient and accurate object detection, representing the 2-stage region-based and 1-stage grid-based approaches, respectively.

Recent research has integrated scene graphs into intelligent systems, enhancing visual cognition by emphasizing relationships and attributes of individual objects[34] and promoting scene understanding[35,36]. Similar to the task cognition work in this paper, scene understanding focuses on extracting target objects and their relational information within background images[37-42]. Liu *et al.* proposed a region-aware attention learning method that prioritizes fine-grained visual regions to improve scene graph generation[37]. Zellers *et al.* introduced a new baseline method for scene graph generation with an architecture to capture substructures[38]. Zhang *et al.* integrated cooking logic into ingredient recognition in food images, generating

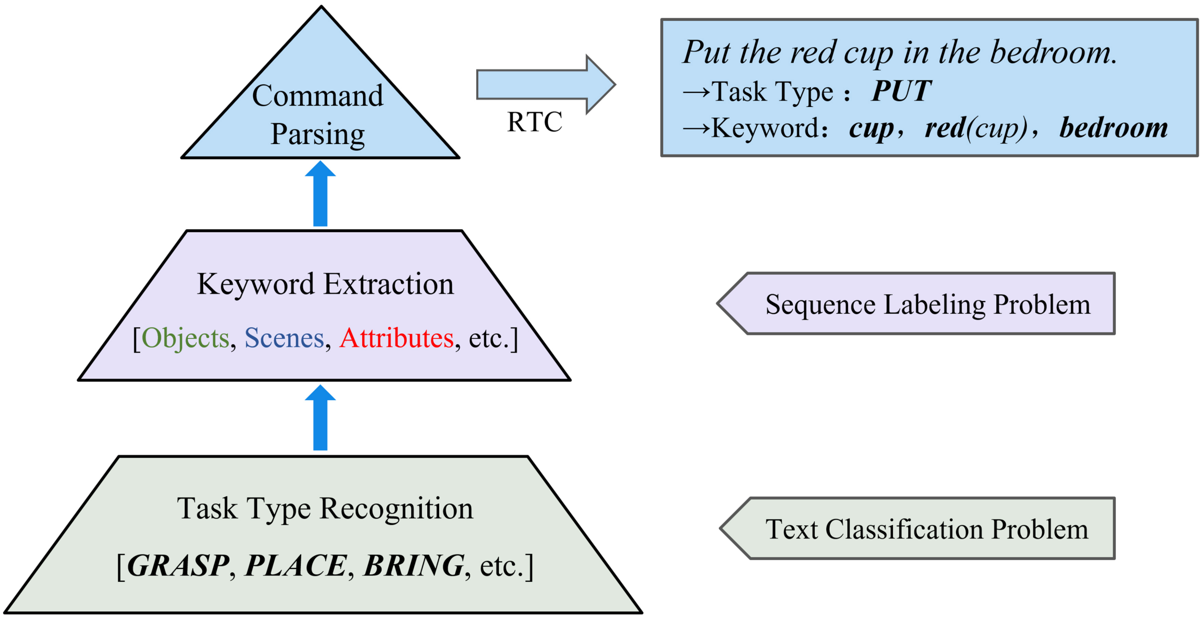

Command Parsing
RTC
Put the red cup in the bedroom.
→Task Type： ***PUT***
→Keyword： ***cup***， ***red****(cup)*， ***bedroom***
Keyword Extraction
[Objects, Scenes, Attributes, etc.]
Sequence Labeling Problem
Task Type Recognition
[***GRASP***, ***PLACE***, ***BRING***, etc.]
Text Classification Problem

Figure 2. Technical roadmap for RTC based on command parsing. For task command parsing, both task type recognition and keyword extraction are required to effectively obtain task information. RTC: Reactive task cognition.

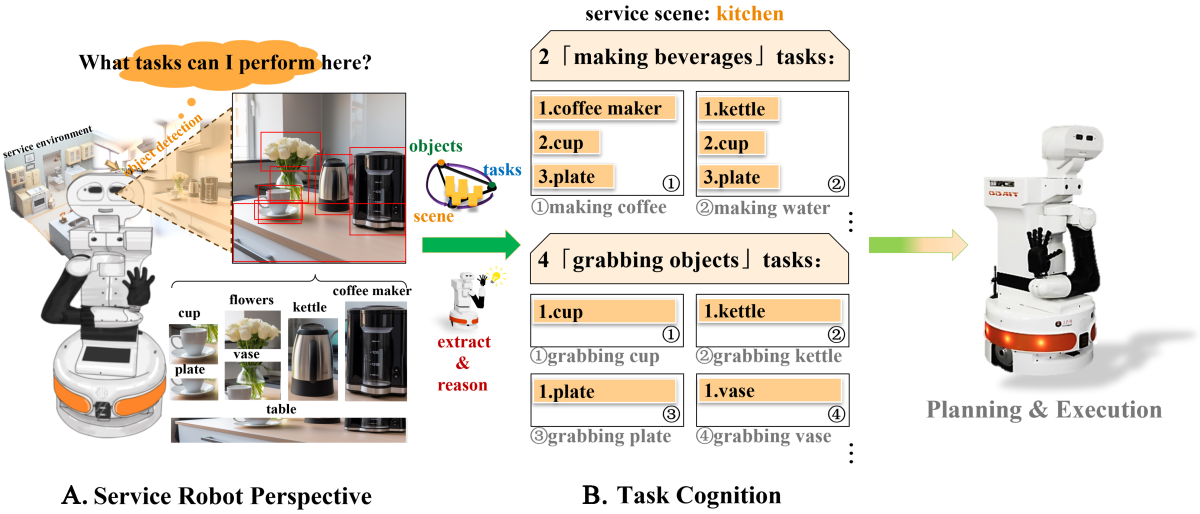


Figure 3. The process of ATC. (A) Using its visual system, the service robot can detect objects (e.g., a coffee maker, a cup, *etc.*) in the image; (B) By extracting relationships among objects and the environment, the robot can reason about executable and suitable task information (including service scene, objects support, and task category) from the current perspective. The information is used for task planning and execution. ATC: Active task cognition.

recipes from extracted semantic information[39]. Zhou *et al.* developed an object search framework combining navigation maps, semantic maps, and scene graphs, enabling robots to track object locations and enhancing autonomous searching and environmental understanding[40]. Riaz *et al.* enhanced scene understanding with scene graphs, equipping robots with semantic knowledge for safer human-machine collaboration (HRC)[41]. Jiao *et al.* introduced a contact graph based on scene graph representation for streamlined robotic operation planning[42]. Both models effectively capture spatial and semantic relationships among objects, highlighting their utility in feature extraction.

Additionally, visual question answering (VQA)[43] is a cross-modal cognitive task that combines CV and NLP to enable machines to answer questions about image content. Recent notable VQA methods[44-47] focus on object relationship reasoning based on images and videos, similar to ATC in robots. Using VQA techniques enhances semantic information extraction[48-50]. Kenfack *et al.* proposed a visual architecture for robotics that processes RGB/RGBD images in real-time to detect objects and calculate their spatial relationships, generating scene or semantic graphs[48]. Das *et al.* introduced a modular strategy learning method for long-range navigation based on language inputs[49]. Luo *et al.* developed a depth and segmentation-based visual attention mechanism to extract local semantic features for VQA using high-speed video segmentation[50]. Several methods have been identified for capturing high-level semantic information, such as attributes and relationships, including the use of external knowledge[51] and graph-based networks[52,53].

With the development of LLMs, notable VQA models such as BLIP-2[54] and Florence[55] have begun incorporating them to enhance multimodal semantic perception. While these methods excel in visual cognition, their focus differs from task cognition, which similarly extracts deep image information - spatial, semantic, and task-related aspects - to identify executable service tasks.

Extracting relational and semantic information from images is essential for visual cognition. Many methods utilize graph neural networks (GNNs), with common structures being graph convolutional networks (GCNs)[56] and graph attention networks (GATs)[57]. Recently, GNN- and GAT-based reasoning has become popular in scene understanding, VQA, and other tasks that analyze relationships between visual elements[58-60].

In graph-structure cognition research, spatial graphs G^{Spa} model the visual features and spatial relationships between detected objects. Each node represents an object I_i , with node features $f_i^{Vis} = \boldsymbol{v}_i$ derived from the visual features. The edges of the spatial graph are described by spatial features f_{ij}^{Spa} derived from bounding box. Given objects I_i and I_j, detected with bounding box features $(x_{i1}, y_{i1}, x_{i2}, y_{i2})$ and $(x_{j1}, y_{j1}, x_{j2}, y_{j2})$, their midpoints are $(x_c^{(i)}, y_c^{(i)})$ and $(x_c^{(j)}, y_c^{(j)})$, and areas are $A^{(i)}$ and $A^{(j)}$. The image area is A, and size is (W, H). The spatial feature $f_{ij}^{Spa} = f_i^{Spa\text{-}rs} \cap f_j^{Spa\text{-}rs} \cap f_{ij}^{Spa\text{-}rp}$ consists of $f_i^{Spa-rs} = \left[\frac{x_{i_1}}{W}, \frac{y_{i_1}}{H}, \frac{x_{i_2}}{W}, \frac{y_{i_2}}{H}, \frac{A^{(i)}}{A}\right]$, $f_j^{Spa-rs} = \left[\frac{x_{j_1}}{W}, \frac{y_{j_1}}{H}, \frac{x_{j_2}}{W}, \frac{y_{j_2}}{H}, \frac{A^{(j)}}{A}\right]$ relative scale features, and relative position feature $f_{ij}^{Spa\text{-}rp}$:

$$f_{ij}^{Spa-rp} = \left[\frac{x_{i_1} - x_{j_1}}{x_{j_2} - x_{j_1}}, \frac{y_{i_1} - y_{j_1}}{y_{j_2} - y_{j_1}}, \log\left(\frac{x_{i_2} - x_{i_1}}{x_{j_2} - x_{j_1}}\right), \log\left(\frac{y_{i_2} - y_{i_1}}{y_{j_2} - y_{j_1}}\right), \frac{x_c^{(i)} - x_c^{(f)}}{W}, \frac{y_c^{(i)} - y_c^{(f)}}{H}\right] \tag{4}$$

Meanwhile, various tailored GNN architectures for specific tasks have been proposed[54-60]. Li *et al.* introduced a simple and interpretable reasoning model that generates visual representations of key scene objects and semantic concepts[60]. This model connects image regions and uses a GCN for reasoning, producing features with semantic relationships and applying a gating and memory mechanism for global reasoning. Yang *et al.* constructed an emotional graph based on semantic concepts and visual features, using GCN to enhance object features with emotional attributes[61]. Liang *et al.* introduced a dual-GAT that integrates semantic, visual, and spatial data, enabling robots to recognize object interactions[62]. Xie *et al.* developed a knowledge-based VQA model that combines visual and non-visual knowledge for question generation[63]. Lyu *et al.* presented a knowledge-enhanced GNN for interpretable recommendations, using an external knowledge base to learn user, object, and interaction representations[64]. Zhang *et al.* and Huang *et al.* employed multi-level reasoning networks to effectively address semantic information loss[65,66]. Many studies utilized GNN structures to capture image features and relationships, enhancing scene understanding

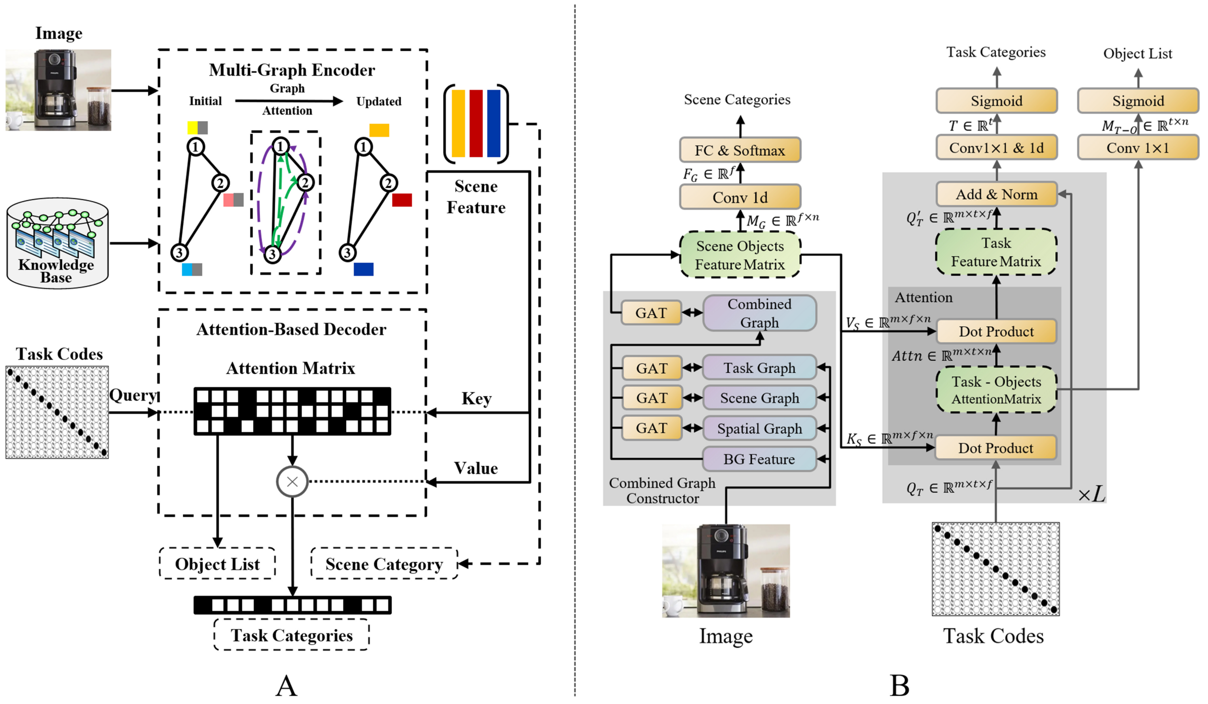


Figure 4. Overall architecture of the ATC-N model[67]. (A) and its components (B). The model features a fused graph made up of three GAT structures: task graph, spatial graph, and semantic graph. This design maximizes the extraction of service scene features and includes a multi-class decoder for predicting task types, corresponding items, and the scene, enabling ATC. ATC-N: Active task cognition network; GAT: graph attention network.

and achieving relevant results. GNNs have proven effective for object relationship handling, particularly in robotic reasoning. Cui *et al.* proposed an active task cognition network (ATC-N) that uses a GNN model for relationship extraction and reasoning, allowing robots to autonomously identify service tasks[67]. The proposed GNN framework for ATC is illustrated in Figure 4.

Below, we provide a comprehensive summary [Table 1] that organizes and compares the task cognition-related literature, highlighting the strengths, weaknesses, and key contributions of various studies in the field. By categorizing and analyzing these studies, we aim to better understand the diverse methodologies, approaches, and technologies proposed and explored. This summary will not only assist researchers in grasping the current state of task cognition but also guide them in identifying gaps, opportunities, and potential areas for future exploration.

In summary, ATC based on visual images is a visual reasoning process that infers task types, related items, and scenes from extracted features and their relationships. Studies show that vision-based relational detection models effectively capture scene features and use attention mechanisms to focus on task objectives. In complex home environments, robots need mechanisms to autonomously discover and understand tasks, establishing dynamic relationships between tasks, items, and scenes. While graph-based reasoning could enhance the analysis of complex relationships, current VQA and visual relation detection models often overlook robotic task cognition, leading to a lack of relevant models and datasets for task information collection. This limits the ability of robots to adapt to environmental changes or predict unknown tasks beyond their predefined scope. Thus, developing an effective reasoning model for robot visual input to facilitate ATC based on visual perception is an urgent challenge.

Table 1. Summary of the advantages and disadvantages of typical task cognition methods applicable to service robots, and key highlights by different categories

	Paper	Year	Advantages of service task cognition	Disadvantages of service task cognition	Highlights
Methods applicable to RTC (based on instruction text)	Abro *et al.*[23]	2022	Significantly reduces the reliance on large annotated datasets and improves the robot's cognitive performance in low-resource environments	The model's complexity and reliance on domain rules limit its scalability and adaptability in service tasks	By incorporating regular expression rules to encode domain knowledge, the model effectively reduces reliance on large supervised datasets and excels in intent detection and slot filling tasks
	Qin *et al.*[24]	2021	The model enhances task interaction through bidirectional information flow, improving the accuracy of the service robot's cognition	The model's high computational complexity and data requirements may limit its use in resource-constrained scenarios	The model captures bidirectional interactions between intent detection and slot filling, significantly improving the performance of the speech language understanding system
	Cheng *et al.*[25]	2021	The method boosts slot label generation performance and inference speed, making it ideal for real-time dialogue systems	Despite faster inference, the model requires extensive training data and resources for complex tasks	The method addresses the slot inconsistency issue caused by the lack of sequential dependencies, improving accuracy and efficiency while maintaining fast inference speed
	He *et al.*[26]	2021	By integrating a knowledge base and shared features, the model excels in intent detection and slot filling accuracy, enhancing the semantic understanding of service robots	The model's complexity increases training and inference costs, particularly in scenarios requiring extensive external knowledge, which may demand higher computational resources	A multi-task intent detection system based on a knowledge base and slot filling model is designed, enhancing semantic understanding and achieving state-of-the-art performance on multiple datasets
	Rajapaksha *et al.*[27]	2020	By applying ontology, the robot can accurately understand and execute unknown terms in high-level commands, enhancing its adaptability and intelligence in complex environments	As the number of unknown terms in commands increases, processing time also rises, which may impact the system's real-time responsiveness, especially under high-load conditions	By leveraging semantic networks and ontology techniques, the system helps robots understand unknown terms in high-level commands, enhancing task execution accuracy at the cognitive level
	Chen *et al.*[51]	2022	Introducing a knowledge graph enhances the robot's ability to understand complex problems, thereby improving task cognition accuracy	The construction and injection of the knowledge graph are complex, potentially requiring significant computational resources and time	A VQA method based on external knowledge graphs is proposed, which converts knowledge into text and effectively integrates it through a delayed injection mechanism, treating the VQA task as a text generation task
	Li *et al.*[54]	2023	By combining frozen models and lightweight modules, the efficiency of visual-language tasks is improved, making it suitable for resource-constrained service robots	Despite saving training parameters, the cost of the pre-training phase remains high, which may pose a challenge for resource-constrained devices	BLIP-2 introduces an efficient pre-training strategy, freezing the image encoder and language model, and using a lightweight query transformer, achieving top performance with fewer parameters
	Xiao *et al.*[55]	2024	Unified text prompts and multi-task learning enhance the robot's adaptability and efficiency in complex tasks	The need for large-scale annotated data and the complexity of model training may increase computational and data resource consumption	A multi-task vision foundation model is designed to perform various visual tasks using text prompts, showcasing strong zero-shot and fine-tuning capabilities, supported by the large-scale annotated dataset FLD-5B
	Mo *et al.*[59]	2021	Graph learning methods effectively integrate multimodal information, helping service robots improve their ability to analyze complex medical images	The network structure is complex, and the graph-based attention module has high computational time complexity	This paper presents a graph learning-based multimodal MRI segmentation method, improving information fusion through graph convolution and attention modules, with excellent results on multiple datasets
	Lyu *et al.*[64]	2022	Enhanced semantic understanding and user behavior inference improve the robot's interpretability and accuracy in recommendation systems	The introduction of external knowledge components may increase the system's resource consumption and computation time	The model integrates external knowledge into the user behavior graph, improving recommendation accuracy, interpretability, and generating human-like recommendation explanations
	Huang *et al.*[66]	2023	It effectively handles occluded images, improving the service robot's cognitive accuracy in complex environments	Although it greatly aids task cognition, the model's high complexity makes it difficult to adapt well to robotic systems	By jointly reasoning image features and compensating for occlusion, the model effectively matches features while suppressing noise and transferring missing semantic information

	Cui *et al.*[67]	2024	A dataset and network structure are specifically designed for robotic task cognition, enabling multi-output classification of task types, scenes, and objects	Although the parameter size is small, the GAT-based network has a long computation time	By using a multi-graph fusion encoder and a multi-task scene understanding decoder, the model effectively captures semantic features of tasks, objects, and scenes, enhancing the robot's intelligence in service tasks

RTC: Reactive task cognition; VQA: visual question answering; BLIP: bootstrapping language-image pre-training; FLD: florence large dataset; MRI: magnetic resonance imaging; GAT: graph attention network.

3. METHODS OF TASK PLANNING IN ROBOTS

In the field of robotics, task planning involves generating an effective sequence of actions for a robot to complete tasks based on specific objectives and environmental conditions. This section provides an overview of the current state of research on various main task planning methods, including classical task planning in robotics, task planning based on LLMs, task planning using scene graphs, and task planning based on reinforcement learning (RL).

3.1. Classic task planning

Classical task planning involves designing planners that generate robotic action strategies, using specialized languages such as planning domain definition language (PDDL)[68] and action language BC[69] to meet diverse planning needs. Robots may need to plan with incomplete information; for instance, Khandelwal *et al.* used BC language to consider action costs in their planning process[70].

Semantic planning for service robots generates sequences of actions based on semantic knowledge suitable for specific scenarios. Savage *et al.* used expert systems to handle incomplete information in planning, while Wang *et al.* enhanced task planning by integrating probabilistic reasoning models, providing item location information[71,72]. Another approach by Wang *et al.* involved pre-training object-level representations to improve planning accuracy and efficiency by understanding environmental dynamics[73].

In handling uncertain task planning, a novel approach[74] combined classical and belief space planning strengths to better manage uncertainty and enhance planning efficiency. To improve interaction during tasks, Bustamante *et al.* introduced constraint action template (CAT) for flexible action adjustments by users, enhancing system usability and user experience[75]. Lastly, Moshayedi *et al.* optimized the FOODIEBOT food delivery robot with classical path planning, achieving the best accuracy with beetle antennae search (BAS) and the highest speed with particle swarm optimization (PSO), while simulation and real-world results closely matched[3].

In industrial robotics, Adu-Bredu *et al.* utilized mixed integer programming (MIP) to optimize specific target functions while meeting all constraints, generating the best task sequences for robots to efficiently manage task allocation and routing in complex environments[76]. Wang *et al.* introduced a new planning method for assembly tasks, allowing robots to autonomously navigate complex industrial settings by leveraging environmental constraints and causal reasoning, resulting in more precise and efficient task execution[77]. Bernardo *et al.* integrated domain ontologies with task planning frameworks to transform agent goals into actionable steps in real-time[78].

In multi-agent systems, goal-directed planning based on module interrelations[79] was proposed, which defines module states and autonomously determines dependencies, enhancing task planning flexibility without manual adjustments. The researcher focused on near-optimal solutions using neighborhood search techniques, accelerating the identification of efficient solutions and reducing computational demands[80]. Task planning for autonomous systems[81] merges offline and online operations, using precomputed data to dynamically generate decisions, thus adapting to changing conditions more robustly. An adaptive robotic task planning framework[82] incorporates user preferences to better coordinate tasks and increase efficiency in human-robot collaborations. Berger *et al.* applied "active queries" to generate 3D models for operational environments, facilitating detailed multi-agent planning based on these models[83].

Overall, classical task planning in industrial settings relies on detailed models and environmental data, optimizing for efficiency but often lacking flexibility. These strategies are widely utilized across various domains[84], enhancing success rates and operational efficiency. However, adapting classical methods to accommodate dynamic and complex home environments remains a challenge that requires further innovation[85].

3.2. LLMs-based task planning

Researchers are enhancing robotic task planning with LLMs to improve adaptability and accuracy. Singh *et al.* used LLMs for more precise robotic arm task execution in desktop environments, enhancing decision-making through real-time environmental feedback [Figure 5][86]. This method improves task success and efficiency.

Additionally, integrating LLMs with advanced visual technologies helps robots adapt to complex environments by providing detailed environmental insights, significantly boosting task outcomes and user experience[87]. Ding *et al.* explored using LLMs with action knowledge bases for dynamic adaptation in uncertain environments, greatly enhancing robotic robustness[88]. Another approach[89] translates natural language into formal task specifications, allowing robots to execute complex tasks accurately and adjust actions in real-time, increasing efficiency and success rates.

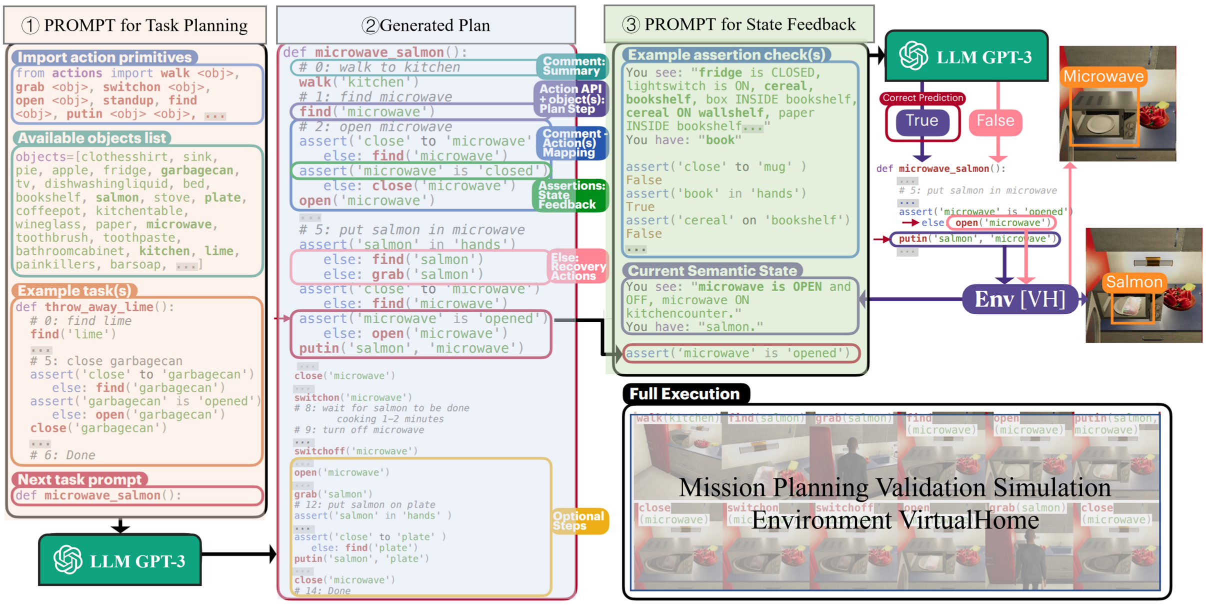
Sarch *et al.* introduced HELPER, an intelligent agent that uses memory-enhanced LLMs to parse and dynamically adjust task plans from human instructions, ideal for home assistant robots interpreting complex tasks from dialogue[90]. This method has proven effective in the TEACh benchmark, enhancing task flexibility and user satisfaction by leveraging dialogue history. Lin *et al.* demonstrated the effectiveness of LLMs in simulation environments, where the models generate actionable plans from high-level descriptions and environmental data, significantly boosting the efficiency of routine tasks[91].

A new benchmarking method[92] involving multimodal LLMs was introduced, combining task progress, visual observations, and language instructions to test the models' ability to handle complex tasks from real-world videos, providing valuable data for optimizing task planning models.

LLM-based task planning methods significantly enhance the environmental understanding and adaptability of robots without needing detailed prior knowledge, making them suitable for dynamic tasks in home and medical settings. These methods leverage strong NLP and environmental perception capabilities of LLMs, allowing robots to execute tasks more flexibly and accurately in complex environments.

3.3. Scene graph-based task planning

Recent years have seen the widespread application of scene graphs in robotic task planning, significantly



```
from actions import walk <obj>,
grab <obj>, switchon <obj>,
open <obj>, standup, find
<obj>, putin <obj> <obj>, ...
```

```
objects=[clothesshirt, sink,
pie, apple, fridge, garbagecan,
tv, dishwashingliquid, bed,
bookshelf, salmon, stove, plate,
coffeepot, kitchentable,
wineglass, paper, microwave,
toothbrush, toothpaste,
bathroomcabinet, kitchen, lime,
painkillers, barsoap, ...]
```

```
def throw_away_lime():
  # 0: find lime
  find('lime')
  ...
  # 5: close garbagecan
  assert('close' to 'garbagecan')
     else: find('garbagecan')
  assert('garbagecan' is 'opened')
     else: open('garbagecan')
  close('garbagecan')
  ...
  # 6: Done
```

```
def microwave_salmon():
```

```
def microwave_salmon():
  # 0: walk to kitchen
  walk('kitchen')
  # 1: find microwave
  find('microwave')
  # 2: open microwave
  assert('close' to 'microwave')
     else: find('microwave')
  assert('microwave' is 'closed')
     else: close('microwave')
  open('microwave')
  ...
  # 5: put salmon in microwave
  assert('salmon' in 'hands')
     else: find('salmon')
     else: grab('salmon')
  assert('close' to 'microwave')
     else: find('microwave')
  assert('microwave' is 'opened')
     else: open('microwave')
  putin('salmon', 'microwave')
  ...
  close('microwave')
  ...
  switchon('microwave')
  # 8: wait for salmon to be done
       cooking 1-2 minutes
  # 9: turn off microwave
  ...
  switchoff('microwave')
  ...
  open('microwave')
  ...
  grab('salmon')
  # 12: put salmon on plate
  assert('salmon' in 'hands' )
  ...
  assert('close' to 'plate' )
     else: find('plate')
  putin('salmon', 'plate')
  ...
  close('microwave')
  # 14: Done
```

```
You see: "fridge is CLOSED,
lightswitch is ON, cereal,
bookshelf, box INSIDE bookshelf,
cereal ON wallshelf, paper
INSIDE bookshelf..."
You have: "book"

assert('close' to 'mug' )
False
assert('book' in 'hands')
True
assert('cereal' on 'bookshelf')
False
...
```

```
You see: "microwave is OPEN and
OFF, microwave ON
kitchencounter."
You have: "salmon."
```

```
assert('microwave' is 'opened')
```

```
def microwave_salmon():
  ...
  # 5: put salmon in microwave
  ...
  assert('microwave' is 'opened')
  else: open('microwave')
  putin('salmon', 'microwave')
  ...
```

Figure 5. Progprompt[86] utilized GPT-3 to guide a robotic arm in performing grasping tasks, with the process validated in VirtualHome simulation environment. The process of using large models is innovatively divided into three modules: ① prompt for planning, ② generated plan, ③ prompt for state feedback.

enhancing planning accuracy and adaptability. Chalvatzaki *et al.* fine-tuned language models with scene graphs to convert complex environmental and task information into structured contexts for detailed robot task planning[93]. Rana *et al.* developed SayPlan, using 3D scene graphs and LLMs to generate high-level task plans and adjust plans iteratively with feedback, creating executable strategies[94]. Agia *et al.* introduced TASKOGRAPHY, evaluating the use of 3D scene graphs in robotic planning by assessing efficiency and exploring real-time planning possibilities[95]. Immorlica *et al.* used a graph-theoretical framework to optimize task planning by modeling tasks as graph nodes and applying graph algorithms[96]. Chen *et al.* utilized a knowledge and-or graph (AOG) with an LSTM network to produce effective atomic action sequences, streamlining robotic operations and enhancing decision-making[97]. These approaches show how integrating advanced modeling techniques such as scene graphs can revolutionize robotic task planning.

Task planning in deterministic and concurrent domains is increasingly focused on efficiency. Kortik *et al.* developed LinGraph, a graph-based planner that streamlines proofs by minimizing irrelevant permutations, optimizing performance in settings with numerous similar objects, such as factories[98]. Sellers *et al.* proposed a safety-aware multi-waypoint navigation and mapping method using the generalized Voronoi diagram and adjacent node selection algorithm, enhancing robot navigation with a local navigator and B-spline smoothing[99]. Meanwhile, Odense *et al.* leveraged GNNs to model planning problems, predicting the runtime of motion planning algorithms to speed up task execution in navigational and manipulative tasks[100]. These methods highlight the role of advanced modeling in improving planning strategies in complex environments.

Facing indeterminate task environments, Kan *et al.* developed a task planning method using stochastic aisle graphs (SAG) that represents task priorities and cost uncertainties[101]. This graph-based approach optimizes task selection and timing within resource constraints, boosting efficiency and effectiveness. Mirakhor *et al.* utilized a directed space graph to enhance item rearrangement processes, leveraging graph embeddings and convolutional networks to enable deep RL planners to efficiently optimize object movement and paths[102].

Saucedo *et al.* introduced belief scene graphs to enhance task planning, particularly for high-level reasoning in scenarios with incomplete information, by integrating expected values into traditional 3D scene graphs[103]. This allows for better prediction and planning in limited information settings, enhancing both scene understanding and the planning of complex tasks such as search and navigation.

Task planning methods using scene graphs offer a structured approach for decomposing and executing complex tasks by utilizing detailed environmental modeling. These methods excel in flexibility and accuracy, particularly in complex scenarios requiring structured task analysis. By leveraging the hierarchical and structured nature of scene graphs, robots can better understand and adapt to complex environments, enhancing task execution success and efficiency. Overall, these approaches showcase significant potential for advancing robotic task planning and adaptability, opening new avenues for intelligent robotics development.

3.4. RL-based task planning

Recent advances in integrating RL with task planning have shown promise. Souza *et al.* applied deep RL using neural networks to enable agents to efficiently navigate complex simulated environments[104]. This method improves adaptability and task performance through continuous learning and interaction with the environment. Liu *et al.* combined PDDL with RL to innovate task and motion planning in complex settings[105]. They developed a vision-based RL actuator that resolves inter-object conflicts with non-grasping actions, enhancing task feasibility. They also introduced a novel reward system that guides the actuator to avoid irreversible failures, significantly improving planning success. The same team explored optimistic RL [Figure 6] to better integrate RL into task planning[106], strengthening the ability to manage uncertainties. These developments highlight the role of RL in enhancing task planning by adapting to complex, uncertain scenarios.

In multi-robot task planning, Li *et al.* applied multi-agent RL (MARL) to streamline task execution for multi-arm robots in orchards, using centralized strategies to reduce conflicts[107]. Wete *et al.* employed safety-focused RL to enhance motion and task planning in automotive manufacturing, integrating safety and compliance checks directly into the planning process[108]. Liu *et al.* developed a hierarchical RL strategy for dynamic mobile robot task planning, improving efficiency by adapting to human movement data[109]. Li *et al.* introduced a multi-vehicle pursuit strategy using adversarial RL, tailored for urban traffic management[110].

For multi-objective planning, Li *et al.* enhanced the efficiency of networked radar systems with RL, optimizing task distribution and resource allocation[111]. Zhang *et al.* created a method for task planning with multiple unmanned surface vessels using advanced deep RL[112]. They segmented the problem into task assignment and collision avoidance, utilizing tailored state and action spaces with matching reward functions, processed via deep neural networks. This approach improved the algorithm's speed and efficiency using enhanced temporal difference methods and a hierarchical system.

The above methods demonstrate the feasibility of using RL for robotic task-level reasoning and planning. The potential of RL to enhance robotic task planning flexibility and performance is significant, guiding robots through complex environments with unique reward mechanisms to handle diverse tasks and adjust strategies dynamically, a crucial trait for domestic service robots.

3.5. Comparison and summary of four types of task planning methods

To summarize the features of each task planning method, Table 2 summarizes the features of four types of

Table 2. Comparison table of features for four types of task planning methods

Planning method	Efficiency	Prior knowledge required	Advantages and disadvantages	Application scenarios
Classical task planning	High	Detailed task models and environmental information	Low versatility, poor adaptability, strong logical sequence	Industrial settings, home services, multi-agent systems
LLM-based task planning	Moderate	No detailed prior knowledge, relies on large datasets	High versatility, stable action sequence generation, suitable for specialized environments	Home environments, medical services, dynamic tasks
Scene graph-based planning	Moderate	Detailed environmental modeling and scene graph data	Strong environmental understanding, adapts to complex tasks through structured decomposition	Environments requiring structured modeling, complex task decomposition
RL-based planning	Low	Minimal or no prior knowledge, relies on real-time feedback learning	High adaptability, suitable for long-term planning, autonomously adapts to changes	Complex dynamic environments, advanced interaction, adaptive control

LLM: Large language model; RL: reinforcement learning.

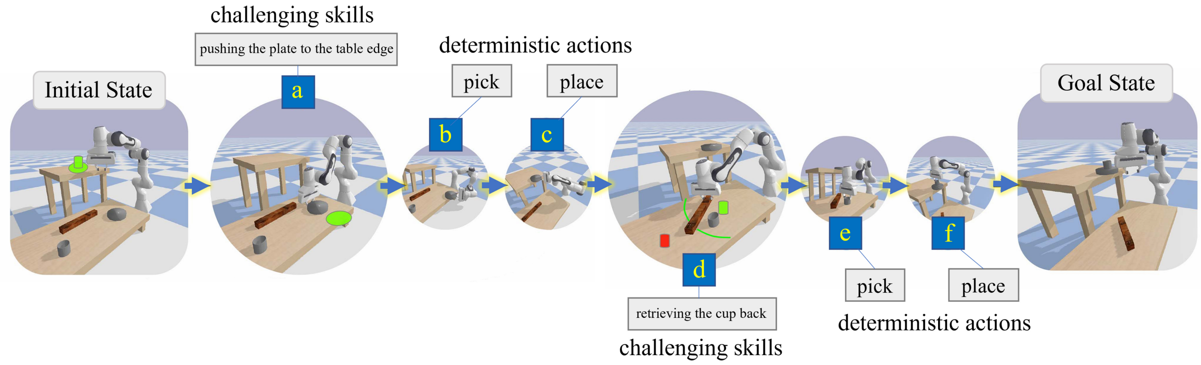


Figure 6. Schematic of optimistic RL-based skill insertions for task and motion planning[106]. With steps a through f illustrating the execution process. It details the identification of initial states and goal states, ultimately demonstrating the correct application of an agent's uncertainty handling skills in grasping tasks. RL: Reinforcement learning.

methods, highlighting their computational efficiency, prior knowledge needs, flexibility, adaptability, and applicable scenarios. It outlines how each method has evolved from classical to modern machine learning-based approaches and their effectiveness in different environments. Each method suits specific applications, and choice depends on the particular needs of the application. In home environments, where efficiency and logical consistency are crucial, classical planning offers high computational efficiency but limited versatility. LLM-based planning compensates for this by providing greater flexibility. A hybrid approach could efficiently generate versatile, executable action sequences, enhancing planning by incorporating external knowledge to improve logical coherence.

To better reflect the characteristics of research related to robot task planning, we have also categorized and summarized the advantages, disadvantages, and highlights of typical studies on service robots [Table 3].

The diversity and complexity of home service tasks necessitate a robot action planning method adaptable to varying environments. Integrating scene graph planning, which requires structured environmental modeling, could address dynamic interactions between items in a home setting. Yet, existing scene models often fall short in representing complex relationships within these environments. To align robot actions with task goals effectively, combining RL-based planning with environmental modeling could optimize adaptability and execution. A hybrid approach could merge the logic of large model-driven planning with the adaptability of RL, maximizing the intelligence and effectiveness of robot task planning in home

Table 3. Summary of the advantages and disadvantages of typical task planning methods applicable to service robots, and key highlights by different categories

	Paper	Year	Advantages of task planning	Disadvantages of task planning	Highlights
Classic task planning	Wang *et al.*[72]	2020	Semantic knowledge and probabilistic reasoning enhance the robot's task planning and adaptability in complex home environments	The hierarchical task network and re-planning mechanism may increase system complexity and task execution delay	A hierarchical task network method based on semantic knowledge and probabilistic reasoning improves task planning and adaptability in uncertain environments for home service robots
	Wang *et al.*[73]	2022	By using synthesized object representations, the robot's task planning and execution capabilities in new environments are improved	Synthetic datasets may differ from real-world environments, limiting the model's generalization ability	A planning method using a synthetic scene dataset is proposed, improving task success rates in multi-step operations by generalizing new object instances with object-level representations
	Adu-Bredu *et al.*[76]	2022	The method generates optimal task plans that meet constraints, improving the robot's task execution efficiency	The method relies on a mixed-integer programming solver, which may encounter computational bottlenecks when handling complex tasks	A robot task planning method encodes the problem as a mixed-integer convex program, using a solver to generate an optimal humanoid robot action sequence that meets all numerical constraints
LLMs-based task planning	Singh *et al.*[86]	2023	The use of a programmatic prompt structure enhances the robot's task planning ability in complex environments	The generated plan may be limited by the robot's current capabilities and may include actions that are not applicable	A programmatic LLM prompt structure generates executable plans across environments, robot capabilities, and tasks, producing action sequences that meet task requirements
	Sarch *et al.*[90]	2023	The method customizes task execution based on user language and habits, enhancing robot adaptability	It relies on an external memory store and continuous updates, with memory management potentially posing challenges	Converting human-robot dialogue into action programs, HELPER, an embodied agent, uses external memory and retrieval-augmented LLM prompts, improving task execution and dialogue performance
	Lin *et al.*[91]	2023	Generating executable plans from high-level goals and environmental data enhances the robot's task planning ability	It relies on environmental data tables, which may pose flexibility and adaptability issues in dynamic environments	Combining high-level goals and environmental data tables, the model generates executable robot plans, enhancing performance through data table encoding, iterative decoding, and evaluation metrics
Scene graph-based task planning	Odense *et al.*[100]	2022	Leveraging GNN to capture problem structure, the efficiency of motion planning and task execution ability is enhanced	The model relies on structured data, which may pose challenges in complex, high-degree-of-freedom tasks	By using GNNs, the model connects motion planning geometry with algorithm runtime, accelerating online planning and identifying subproblems suited for specific algorithms
	Kan *et al.*[101]	2021	The method efficiently optimizes task planning in uncertain environments, enhancing resource utilization	The method may face challenges in real-time decision-making and execution when dealing with complex or highly dynamic environments	Addressing cost uncertainty in precision agriculture task planning, the NBA-P algorithm based on SAG outperforms other allocation methods
	Mirakhor *et al.*[102]	2024	It improves the efficiency of object rearrangement in multi-room environments and can handle unseen objects and obstacles	The method's complexity may lead to computational challenges in certain complex environments	A novel task planning method is proposed for discovering unseen objects and rearranging in multi-room environments, effectively reducing travel paths using deep RL and spatial graph techniques
RL-based task planning	Liu *et al.*[106]	2024	By integrating RL skills, the robot's task planning capability and efficiency in uncertain environments are improved	Dependence on uncertainty handling may lead to adaptability issues in highly dynamic scenarios	The method integrating RL skills into the TAMP pipeline, using data-driven logic for symbolic planning and optimization to address uncertainty
	Li *et al.*[107]	2023	Task planning efficiency is improved and operational collaboration is optimized for multi-arm picking robots	Training the model may require substantial computational resources and time, and could face challenges in dynamic environments	A task planning strategy for a four-arm picking robot is proposed, using a Markov game framework to avoid the computational complexity of NP-hard scheduling problems, and trained through MARL structure
	Liu *et al.*[109]	2023	The robot's task planning efficiency and environmental adaptability in dynamic environments are effectively enhanced by its HRL approach	The combination of symbolic planning and HRL may increase system complexity and pose significant real-time challenges in highly dynamic environments	A novel HA-GHDQ algorithm combines symbolic planning and HRL, using human motion patterns (MoDs) in dynamic environments to generate long-term task planning strategies for robots

LLMs: Large language models; GNN: graph neural network; NBA-P: next-best-action planning; SAG: stochastic aisle graph; RL: reinforcement learning; TAMP: task and motion planning; NP: non-deterministic polynomial time; MARL: multi-agent reinforcement learning; HRL: hierarchical reinforcement learning; MoDs: motion dynamics.

services. This would ensure that robots can perform diverse tasks intelligently and efficiently.

4. CURRENT ISSUES

From the review and analysis, it is evident that significant efforts have been made globally to enhance the intelligence of domestic robots, achieving some stage-wise results. However, due to the complexity and diversity of the home service environment, effective cognition and task planning by home robots, followed by smooth execution of service tasks, still require extensive exploration. Currently, the primary issues in the cognition and planning field for service robots include the following areas:

(1) Challenges in language interpretation and task constraint analysis: Robots struggle with interpreting natural language instructions due to varied user habits and traditional methods reliant on syntactic analysis, which often result in errors and lack depth for complex commands. Additionally, research lacks in analyzing task constraints, affecting the accuracy of understanding and planning tasks.

(2) Lack of active cognition and adaptive capabilities in robots: Robots lack active cognition abilities, primarily responding reactively to commands which limits their operational potential in complex environments. The absence of effective autonomous discovery and cognition mechanisms hinders establishing dynamic relationships necessary for adapting to environmental changes or anticipating unknown tasks.

(3) Limitations of rigid action planning and inadequate task flexibility: Current action sequence planning often depends on rigid templates, limiting its ability to adapt to new tasks and understand specific service requirements, which results in action sequences that fail to meet practical demands. These methods also lack the use of detailed prior knowledge, leading to sequences that are impractical for real-world domestic tasks. This results in inefficient and illogical task execution.

(4) Limitations of task planning in dynamic environments: Traditional task planning struggles in dynamic, uncertain environments due to reliance on deterministic models that cannot account for real-time changes. This results in static action sequences unfit for the current conditions. Additionally, the models fail to account for interactions between dynamic and static items in home settings, lacking comprehensive scene modeling.

(5) Hybrid approaches are needed to integrate planning algorithms' strengths: Classical and LLM-based planning methods provide stability but lack adaptability; scene graph and RL methods are responsive but resource-intensive and prone to delays. A hybrid approach could improve adaptability, necessitating a platform that integrates both cognition and planning to manage challenges such as resource distribution and system complexity.

5. FUTURE DIRECTIONS

To address existing issues, the proposed research directions for enhancing task cognition and planning in home service robots include:

● Enhancing language models: Related research studies have fully confirmed that enhancing language intelligence parsing models[22,24,26] can improve the ability of robots to interpret user intentions. However, the instruction parsing models specifically for service robots are not yet particularly robust. Developing instruction parsing models using word embeddings can efficiently extract and relate task classifications and keywords. Refining large models will enhance the precision of command interpretation, providing reliable data for robot task planning.

● Inferring environmental information: For image-based cognitive methods, the researchers have provided sufficient feasibility validation[50,59,62,67]. However, current models specifically focused on task cognition have not received much attention. By utilizing visual cognition and graph-based modeling, robots can simulate proactive observation capabilities, thereby improving the accuracy and interpretability of environmental understanding.

● Multimodal task cognition: Many studies have already combined large models with cognition and demonstrated their feasibility[48,51,54,55]. By integrating instructions with visual data, robots can enhance their comprehension of real-time tasks, leveraging the synergy between textual commands and environmental context.

For task planning, future research directions could include:

● Knowledge-guided action sequences: The introduction of knowledge modules to improve task planning accuracy has been demonstrated to enhance the precision of robot planning[72,77,88]. Combine the logical consistency of classical planning methods with the adaptability of LLMs, utilizing prior knowledge to enhance planning reliability and enable autonomous action sequence generation.

● Environment-based action planning: The researchers have demonstrated that environment modeling[98,101] and RL methods[104,106] enable planning results to adapt to dynamic environments, significantly advancing service robot planning research. By leveraging these techniques, robots can anticipate and execute appropriate task actions in dynamic home environments, thereby enhancing their adaptability.

● Integrated task planning: The task cognition and planning of service robots is an integrated process. Design a data processing system that combines cognition and planning through modular design and multi-channel information input, focusing on resource allocation and scheduling to meet the demands of home service robots efficiently.

6. CONCLUSIONS

This review examines current approaches to task cognition and planning for domestic service robots, highlighting key challenges and advances. Due to the dynamic nature of home environments and complex user instructions, traditional rule-based methods struggle to scale, while deep learning shows promise in understanding instructions but faces generalization challenges.

We emphasize the shift from RTC to ATC, where robots not only respond to explicit commands but also actively identify tasks through environmental observation. Multimodal data, such as visual and contextual information, improves task recognition, while combining LLMs with logical planning frameworks (e.g., PDDL) offers new ways to generate reliable action sequences.

Future developments in RL, attention mechanisms, and real-time environment modeling will further enhance planning and execution efficiency. Integrating expert knowledge into task planning will improve robot intelligence and execution. The goal is to create flexible, human-like cognition and planning methods that enhance interaction, reliability, and adaptability in diverse home environments.

DECLARATIONS

Authors' contributions

Funding acquisition: Zhang, Y.; Zhang, C. H.
Writing - original draft: Cui, Y.; Zhang, Y.
Writing - review and editing: Cui, Y., Zhang, Zhang, C. H., Yang, S. X.

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China under Grants 62203378 and 62203377, the Hebei Natural Science Foundation under Grant Nos. F2024203036, F2024203115 and F2022203097, the Science Research Project of Hebei Education Department under Grant BJK2024195, the S&T Program of Hebei under Grant 236Z2002G, 236Z1603G, and the Hebei Innovation Capability Improvement Plan Project under Grant No. 22567619H.

Conflicts of interest

Yang, S. X. is Editor in Chief and Zhang, Y. is Junior Editorial Board Member of the journal *Intelligence & Robotics*. Yang, S. X. and Zhang, Y. were not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, or decision-making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Mukherjee, D.; Gupta, K.; Chang, L. H.; Najjaran, H. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robot. Cim. Int. Manuf.* **2022**, *73*, 102231. DOI
2. Sánchez-Ibáñez, J. R.; Pérez-Del-Pulgar, C. J.; García-Cerezo, A. Path planning for autonomous mobile robots: a review. *Sensors* **2021**, *21*, 7898. DOI PubMed PMC
3. Moshayedi, A. J.; Roy, A. S.; Liao, L.; Khan, A. S.; Kolahdooz, A.; Eftekhari, A. Design and development of FOODIEBOT robot: from simulation to design. *IEEE. Access.* **2024**, *12*, 36148-72. DOI
4. Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.; Shi, P. Deep learning-based scene understanding for autonomous robots: a survey. *Intell. Robot.* **2023**, *3*, 374-401. DOI
5. Zhou, C.; Huang, B.; Fränti, P. A review of motion planning algorithms for intelligent robots. *J. Intell. Manuf.* **2022**, *33*, 387-424. DOI
6. Zhang, Y.; Tian, G.; Shao, X. Safe and efficient robot manipulation: task-oriented environment modeling and object pose estimation. *IEEE. Trans. Instrum. Meas.* **2021**, *70*, 1-12. DOI
7. Church, K. W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155-62. DOI
8. Pennington, J.; Socher, R.; Manning, C. D. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25-29; Doha, Qatar. 2014, pp. 1532-43. Available from: https://aclanthology.org/D14-1162.pdf. (accessed 2025-01-21).
9. Weld, H.; Huang, X.; Long, S.; Poon, J.; Han, S. C. A survey of joint intent detection and slot filling models in natural language understanding. *ACM. Comput. Surv.* **2023**, *55*, 1-38. DOI
10. Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; Daumé, H. I. I. I. Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26-31; Beijing, China. 2015, pp. 1681-91. DOI
11. Jiang, C.; Xu, Q.; Song, Y.; Yuan, X.; Pang, B.; Li, Y. Discrete sequence rearrangement based self-supervised chinese named entity recognition for robot instruction parsing. *Intell. Robot.* **2023**, *3*, 337-54. DOI
12. Kleenankandy, J.; K, A. A. N. An enhanced Tree-LSTM architecture for sentence semantic modeling using *typed dependencies*. *Inform. Process. Manag.* **2020**, *57*, 102362. DOI
13. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE.* **1998**, *86*, 2278-324. DOI
14. Ayetiran, E. F. Attention-based aspect sentiment classification using enhanced learning through cnn-Bilstm networks. *Knowl. Based. Syst.* **2022**, *252*, 109409. DOI
15. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *arXiv* **2024**, arXiv:1706.03762. Available online: https://doi.org/10.48550/arXiv.1706.03762 (accessed 21 Jan 2025)
16. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J Mach Learn Res* 2011;12:2493-537. Available from: https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf?source. (accessed 2025-01-21).
17. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural. Comput.* **2019**, *31*, 1235-70. DOI
18. Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* **2024**, arXiv:1810.04805. Available online: https://doi.org/10.48550/arXiv.1810.04805 (accessed 21 Jan 2025)
19. Achiam, J.; Adler, S.; Agarwal, S.; et al. Gpt-4 technical report. *arXiv* **2024**, arXiv:2303.08774. Available online: https://doi.org/10.48550/arXiv.2303.08774 (accessed 21 Jan 2025)
20. Zhang, C.; Chen, J.; Li, J.; Peng, Y.; Mao, Z. Large language models for human–robot interaction: a review. *Biomim. Intell. Robot.* **2023**, *3*, 100131. DOI
21. Liu, B.; Lane, I. Attention-Based recurrent neural network models for joint intent detection and slot filling. *Interspeech* **2016**, *2016*, 685-9. DOI
22. Goo, C. W.; Gao, G.; Hsu, Y. K.; et al. Slot-gated modeling for joint slot filling and intent prediction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018 Jun 1-6; New Orleans, Louisiana, USA. 2018, pp. 753-7. DOI
23. Abro, W. A.; Qi, G.; Aamir, M.; Ali, Z. Joint intent detection and slot filling using weighted finite state transducer and BERT. *Appl. Intell.* **2022**, *52*, 17356-70. DOI
24. Qin, L.; Liu, T.; Che, W.; Kang, B.; Zhao, S.; Liu, T. A co-interactive transformer for joint slot filling and intent detection. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021 Jun 6-11; Toronto, Canada. IEEE; 2021. pp. 8193-7. DOI
25. Cheng, L.; Jia, W.; Yang, W. An effective non-autoregressive model for spoken language understanding. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; 2021 Nov 1-5; Boise, USA. Association for Computing Machinery; 2021, pp. 241-50. DOI
26. He, T.; Xu, X.; Wu, Y.; Wang, H.; Chen, J. Multitask learning with knowledge base for joint intent detection and slot filling. *Appl. Sci.* **2021**, *11*, 4887. DOI
27. Rajapaksha, U. U. S.; Jayawardena, C. Ontology based optimized algorithms to communicate with a service robot using a user

command with unknown terms. In: 2020 2nd International Conference on Advancements in Computing (ICAC); 2020 Dec 10-11; Malabe, Sri Lanka. IEEE; 2020. pp. 258-62. DOI

28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2017**, *39*, 1137-49. DOI PubMed
29. Duan, H.; Yang, Y.; Li, D.; Wang, P. Human–robot object handover: recent progress and future direction. *Biomim. Intell. Robot.* **2024**, *4*, 100145. DOI
30. Wang, C. Y.; Yeh, I. H.; Liao, H. Y. M. YOLOv9: learning what you want to learn using programmable gradient information. *arXiv* **2024**, arXiv:2402.13616. Available online: https://doi.org/10.48550/arXiv.2402.13616 (accessed 21 Jan 2025)
31. Zhang, Y.; Yin, M.; Wang, H.; Hua, C. Cross-level multi-modal features learning with transformer for RGB-D object recognition. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2023**, *33*, 7121-30. DOI
32. Liu, W.; Anguelov, D.; Erhan, D.; et al. SSD: Single shot multibox detector. In: Proceedings of the Computer Vision–ECCV 2016: 14th European Conference; 2016 October 11-14; Amsterdam, the Netherlands. Springer International Publishing; 2016, pp. 21-37. DOI
33. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11-17; Montreal, Canada. IEEE; 2021. pp. 2778-88. DOI
34. Zhang, Y.; Tian, G.; Chen, H. Exploring the cognitive process for service task in smart home: a robot service mechanism. *Future. Gener. Comput. Syst.* **2020**, *102*, 588-602. DOI
35. Xu, D.; Zhu, Y.; Choy, C. B.; Li, F. F. Scene graph generation by iterative message passing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21-26; Honolulu, USA. IEEE; 2017. pp. 5410-9. DOI
36. Dai, B.; Zhang, Y.; Lin, D. Detecting visual relationships with deep relational networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21-26; Honolulu, USA. IEEE; 2017. pp. 3076-86. DOI
37. Liu, A. A.; Tian, H.; Xu, N.; Nie, W.; Zhang, Y.; Kankanhalli, M. Toward region-aware attention learning for scene graph generation. *IEEE. Trans. Neural. Netw. Learn. Syst.* **2022**, *33*, 7655-66. DOI
38. Zellers, R.; Yatskar, M.; Thomson, S.; Choi, Y. Neural motifs: scene graph parsing with global context. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, USA. IEEE; 2018. pp. 5831-40. DOI
39. Zhang, M.; Tian, G.; Zhang, Y.; Liu, H. Sequential learning for ingredient recognition from images. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2023**, *33*, 2162-75. DOI
40. Zhou, F.; Liu, H.; Zhao, H.; Liang, L. Long-term object search using incremental scene graph updating. *Robotica* **2023**, *41*, 962-75. DOI
41. Riaz, H.; Terra, A.; Raizer, K.; Inam, R.; Hata, A. Scene understanding for safety analysis in human-robot collaborative operations. In: 2020 6th International Conference on Control, Automation and Robotics (ICCAR); 2020 Apr 20-23; Singapore. IEEE; 2020. pp. 722-31. DOI
42. Jiao, Z.; Niu, Y.; Zhang, Z.; Zhu, S. C.; Zhu, Y.; Liu, H. Sequential manipulation planning on scene graph. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23-27; Kyoto, Japan. IEEE; 2022. pp. 8203-10. DOI
43. Antol, S.; Agrawal, A.; Lu, J.; et al. Vqa: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7-13; Santiago, Chile. IEEE; 2015. pp. 2425-33. DOI
44. Li, G.; Wang, X.; Zhu, W. Boosting visual question answering with context-aware knowledge aggregation. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12-16; Melbourne, Australia. Association for Computing Machinery; 2020. pp. 1227-35. DOI
45. Wang, P.; Yang, A.; Men, R.; et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Proceedings of the 39th International Conference on Machine Learning; Baltimore, Maryland. 2022. pp. 23318-40. Available from: https://proceedings.mlr.press/v162/wang22al.html. (accessed 2025-01-21).
46. Lu, P.; Ji, L.; Zhang, W.; Duan, N.; Zhou, M.; Wang, J. R-VQA: learning visual relation facts with semantic attention for visual question answering. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19-23; New York, USA. Association for Computing Machinery; 2018. pp. 1880-9. DOI
47. Yu, T.; Yu, J.; Yu, Z.; Huang, Q.; Tian, Q. Long-term video question answering via multimodal hierarchical memory attentive networks. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2021**, *31*, 931-44. DOI
48. Kenfack, F. K.; Siddiky, F. A.; Balint-Benczedi, F.; Beetz, M. Robotvqa - a scene-graph-and deep-learning-based visual question answering system for robot manipulation. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24 - 2021 Jan 24; Las Vegas, USA. IEEE; 2020. pp. 9667-74. DOI
49. Das, A.; Gkioxari, G.; Lee, S.; Parikh, D.; Batra, D. Neural modular control for embodied question answering. *arXiv* **2024**, arXiv:1810.11181. Available online: https://doi.org/10.48550/arXiv.1810.11181 (accessed 21 Jan 2025)
50. Luo, H.; Lin, G.; Yao, Y.; Liu, F.; Liu, Z.; Tang, Z. Depth and video segmentation based visual attention for embodied question answering. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2023**, *45*, 6807-19. DOI
51. Chen, Z.; Huang, Y.; Chen, J.; et al. LAKO: knowledge-driven visual question answering via late knowledge-to-text injection. *arXiv* **2024**, arXiv:2207.12888. Available online: https://doi.org/10.48550/arXiv.2207.12888 (accessed 21 Jan 2025)
52. Teney, D.; Liu, L.; van, D. H. A. Graph-structured representations for visual question answering. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 July 21-26; Honolulu, USA. IEEE; 2017. pp. 1-9. DOI

53. Norcliffe-Brown, W.; Vafeias, E.; Parisot, S. Learning conditioned graph structures for interpretable visual question answering. *arXiv* **2024**, arXiv:1806.07243. Available online: https://doi.org/10.48550/arXiv.1806.07243 (accessed 21 Jan 2025)
54. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2024**, arXiv:2301.12597. Available online: https://doi.org/10.48550/arXiv.2301.12597 (accessed 21 Jan 2025)
55. Xiao, B.; Wu, H.; Xu, W.; et al. Florence-2: advancing a unified representation for a variety of vision tasks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16-22; Seattle, USA. IEEE; 2024. pp. 4818-29. DOI
56. Bhatti, U. A.; Tang, H.; Wu, G.; Marjan, S.; Hussain, A.; Sarker, S. K. Deep learning with graph convolutional networks: an overview and latest applications in computational intelligence. *Int. J. Intell. Syst.* **2023**, *2023*, 8342104. DOI
57. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2024**, arXiv:1710.10903. Available online: https://doi.org/10.48550/arXiv.1710.10903 (accessed 21 Jan 2025)
58. Chen, Z. M.; Wei, X. S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, USA. IEEE; 2019. pp. 5177-86. DOI
59. Mo, S.; Cai, M.; Lin, L.; et al. Mutual information-based graph co-attention networks for multimodal prior-guided magnetic resonance imaging segmentation. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2022**, *32*, 2512-26. DOI
60. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual semantic reasoning for image-text matching. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27 - Nov 2; Seoul, South Korea. IEEE; 2019. pp. 4654-62. DOI
61. Yang, J.; Gao, X.; Li, L.; Wang, X.; Ding, J. SOLVER: scene-object interrelated visual emotion reasoning network. *IEEE. Trans. Image. Process.* **2021**, *30*, 8686-701. DOI
62. Liang, Z.; Liu, J.; Guan, Y.; Rojas, J. Visual-semantic graph attention networks for human-object interaction detection. In: 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO); 2021 Dec 27-31; Sanya, China. IEEE; 2021. pp. 1441-7. DOI
63. Xie, J.; Fang, W.; Cai, Y.; Huang, Q.; Li, Q. Knowledge-based visual question generation. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2022**, *32*, 7547-58. DOI
64. Lyu, Z.; Wu, Y.; Lai, J.; Yang, M.; Li, C.; Zhou, W. Knowledge enhanced graph neural networks for explainable recommendation. *IEEE. Trans. Knowl. Data. Eng.* **2023**, *35*, 4954-68. DOI
65. Zhang, L.; Wang, S.; Liu, J.; et al. MuL-GRN: multi-level graph relation network for few-shot node classification. *IEEE. Trans. Knowl. Data. Eng.* **2023**, *35*, 6085-98. DOI
66. Huang, M.; Hou, C.; Yang, Q.; Wang, Z. Reasoning and tuning: graph attention network for occluded person re-identification. *IEEE. Trans. Image. Process.* **2023**, *32*, 1568-82. DOI
67. Cui, Y.; Tian, G.; Jiang, Z.; Zhang, M.; Gu, Y.; Wang, Y. An active task cognition method for home service robot using multi-graph attention fusion mechanism. *IEEE. Trans. Circuits. Syst. Video. Technol.* **2024**, *34*, 4957-72. DOI
68. Ghallab, M.; Knoblock, C.; Wilkins, D.; et al. PDDL - The planning domain definition language. Washington: University of Washington Press; 1998. pp. 1-27. Available from: https://www.researchgate.net/publication/2278933_PDDL_-_The_Planning_Domain_Definition_Language. (accessed 2025-01-21).
69. Lee, J.; Lifschitz, V.; Yang, F. Action language BC: preliminary report. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence; Beijing, China, 2013; pp. 983-9. Available from: https://www.ijcai.org/Proceedings/13/Papers/150.pdf. (accessed 2025-01-21).
70. Khandelwal, P.; Yang, F.; Leonetti, M.; Lifschitz, V.; Stone, P. Planning in action language BC while learning action costs for mobile robots. *ICAPS.* **2014**, *24*, 472-80. DOI
71. Savage, J.; Rosenblueth, D. A.; Matamoros, M.; et al. Semantic reasoning in service robots using expert systems. *Robot. Auton. Syst.* **2019**, *114*, 77-92. DOI
72. Wang, Z.; Tian, G.; Shao, X. Home service robot task planning using semantic knowledge and probabilistic inference. *Knowl. Based. Syst.* **2020**, *204*, 106174. DOI
73. Wang, C.; Xu, D.; Li, F. F. Generalizable task planning through representation pretraining. *IEEE. Robot. Autom. Lett.* **2022**, *7*, 8299-306. DOI
74. Adu-Bredu, A.; Zeng, Z.; Pusalkar, N.; Jenkins, O. C. Elephants don't pack groceries: robot task planning for low entropy belief states. *IEEE. Robot. Autom. Lett.* **2022**, *7*, 25-32. DOI
75. Bustamante, S.; Quere, G.; Leidner, D.; Vogel, J.; Stulp, F. CATs: task planning for shared control of assistive robots with variable autonomy. In: 2022 International Conference on Robotics and Automation (ICRA); 2022 May 23-27; Philadelphia, USA. IEEE; 2022. pp. 3775-82. DOI
76. Adu-Bredu, A.; Devraj, N.; Jenkins, O. C. Optimal constrained task planning as mixed integer programming. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23-27; Kyoto, Japan. IEEE; 2022. pp. 12029-36. DOI
77. Wang, Z.; Gan, Y.; Dai, X. Assembly-oriented task sequence planning for a dual-arm robot. *IEEE. Robot. Autom. Lett.* **2022**, *7*, 8455-62. DOI
78. Bernardo, R.; Sousa, J. M.; Gonçalves, P. J. Planning robotic agent actions using semantic knowledge for a home environment. *Intell. Robot.* **2021**, *1*, 101-5. DOI
79. Yano, T.; Ito, K. Goal-oriented task planning for composable robot control system using module-based control-as-inference framework. In: 2024 IEEE/SICE International Symposium on System Integration (SII); 2024 Jan 8-11; Ha Long, Vietnam. IEEE;

2024. pp. 1219-26. DOI

80. Zeng, F.; Shirafuji, S.; Fan, C.; Nishio, M.; Ota, J. Stepwise large-scale multi-agent task planning using neighborhood search. *IEEE. Robot. Autom. Lett.* **2024**, *9*, 111-8. DOI
81. Li, S.; Wei, M.; Li, S.; Yin, X. Temporal logic task planning for autonomous systems with active acquisition of information. *IEEE. Trans. Intell. Veh.* **2024**, *9*, 1436-49. DOI
82. Noormohammadi-Asl, A.; Smith, S. L.; Dautenhahn, K. To lead or to follow? Adaptive robot task planning in human-robot collaboration. *arXiv* **2024**, arXiv:2401.01483. Available online: https://doi.org/10.48550/arXiv.2401.01483 (accessed 21 Jan 2025)
83. Berger, C.; Doherty, P.; Rudol, P.; Wzorek, M. Leveraging active queries in collaborative robotic mission planning. *Intell. Robot.* **2024**, *4*, 87-106. DOI
84. Zhang, Y.; Tian, G.; Shao, X.; Zhang, M.; Liu, S. Semantic grounding for long-term autonomy of mobile robots toward dynamic object search in home environments. *IEEE. Trans. Ind. Electron.* **2023**, *70*, 1655-65. DOI
85. Zhang, Y.; Tian, G.; Shao, X.; Cheng, J. Effective safety strategy for mobile robots based on laser-visual fusion in home environments. *IEEE. Trans. Syst. Man. Cybern. Syst.* **2022**, *52*, 4138-50. DOI
86. Singh, I.; Blukis, V.; Mousavian, A.; et al. Progprompt: generating situated robot task plans using large language models. In: 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29 - Jun 02; London, UK. IEEE; 2023. pp. 11523-30. DOI
87. Wang, L.; Ma, C.; Feng, X.; et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **2024**, *18*, 40231. DOI
88. Ding, Y.; Zhang, X.; Amiri, S.; et al. Integrating action knowledge and LLMs for task planning and situation handling in open worlds. *Auton. Robot.* **2023**, *47*, 981-97. DOI
89. Pallagani, V.; Muppasani, B. C.; Roy, K.; et al. On the prospects of incorporating large language models (LLMs) in automated planning and scheduling (APS). *ICAPS.* **2024**, *34*, 432-44. DOI
90. Sarch, G.; Wu, Y.; Tarr, M. J.; Fragkiadaki, K. Open-ended instructable embodied agents with memory-augmented large language models. *arXiv* **2024**, arXiv:2301.15127. Available online: https://doi.org/10.48550/arXiv.2310.15127 (accessed 21 Jan 2025)
91. Lin, B. Y.; Huang, C.; Liu, Q.; Gu, W.; Sommerer, S.; Ren, X. On grounded planning for embodied tasks with language models. *AAAI.* **2023**, *37*, 13192-200. DOI
92. Akiyama, S.; Dossa, R. F.; Arulkumaran, K.; Sujit, S.; Johns, E. Open-loop VLM robot planning: an investigation of fine-tuning and prompt engineering strategies. In: Proceedings of the First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024; Yokohama, Japan. 2024. pp. 1-6. Available from: https://openreview.net/forum?id=JXngwwPMR5. (accessed 2025-01-21).
93. Chalvatzaki, G.; Younes, A.; Nandha, D.; Le, A. T.; Ribeiro, L. F. R.; Gurevych, I. Learning to reason over scene graphs: a case study of finetuning GPT-2 into a robot language model for grounded task planning. *Front. Robot. AI.* **2023**, *10*, 1221739. DOI PubMed PMC
94. Rana, K.; Haviland, J.; Garg, S.; Abou-Chakra, J.; Reid, I. D.; Suenderhauf, N. SAYPLAN: grounding large language models using 3D scene graphs for scalable robot task planning. *arXiv* **2024**, arXiv:2307.06135. Available online: https://doi.org/10.48550/arXiv.2307.06135 (accessed 21 Jan 2025)
95. Agia, C.; Jatavallabhula, K. M.; Khodeir, M.; et al. TASKOGRAPHY: evaluating robot task planning over large 3D scene graphs. *arXiv* **2024**, arXiv:2207.05006. Available online: https://doi.org/10.48550/arXiv.2207.05006 (accessed 21 Jan 2025)
96. Immorlica, N. Technical perspective: a graph-theoretic framework traces task planning. *Commun. ACM.* **2018**, *61*, 98-98. DOI
97. Chen, T.; Chen, R.; Nie, L.; Luo, X.; Liu, X.; Lin, L. Neural task planning with AND–OR graph representations. *IEEE. Trans. Multimed.* **2019**, *21*, 1022-34. DOI
98. Kortik, S.; Saranli, U. LinGraph: a graph-based automated planner for concurrent task planning based on linear logic. *Appl. Intell.* **2017**, *47*, 914-34. DOI
99. Sellers, T.; Lei, T.; Luo, C.; Jan, G. E.; Junfeng, M. A node selection algorithm to graph-based multi-waypoint optimization navigation and mapping. *Intell. Robot.* **2022**, *2*, 333-54. DOI
100. Odense, S.; Gupta, K.; Macready, W. G. Neural-guided runtime prediction of planners for improved motion and task planning with graph neural networks. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23-27; Kyoto, Japan. IEEE; 2022. pp. 12471-8. DOI
101. Kan, X.; Thayer, T. C.; Carpin, S.; Karydis, K. Task planning on stochastic aisle graphs for precision agriculture. *IEEE. Robot. Autom. Lett.* **2021**, *6*, 3287-94. DOI
102. Mirakhor, K.; Ghosh, S.; Das, D.; Bhowmick, B. Task planning for object rearrangement in multi-room environments. *AAAI.* **2024**, *38*, 10350-7. DOI
103. Saucedo, M. A. V.; Patel, A.; Saradagi, A.; Kanellakis, C.; Nikolakopoulos, G. Belief scene graphs: expanding partial scenes with objects through computation of expectation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13-17; Yokohama, Japan. IEEE; 2024. pp. 9441-7. DOI
104. Souza, C.; Velhor, L. Deep reinforcement learning for task planning of virtual characters. *Intell. Comput.* **2021**, *284*, 694-711. DOI
105. Liu, G.; de, W. J.; Steckelmacher, D.; Hota, R. K.; Nowe, A.; Vanderborght, B. Synergistic task and motion planning with reinforcement learning-based non-prehensile actions. *IEEE. Robot. Autom. Lett.* **2023**, *8*, 2764-71. DOI
106. Liu, G.; de, W. J.; Durodié, Y.; Steckelmacher, D.; Nowe, A.; Vanderborght, B. Optimistic reinforcement learning-based skill

insertions for task and motion planning. *IEEE. Robot. Autom. Lett.* **2024**, *9*, 5974-81. DOI

107. Li, T.; Xie, F.; Qiu, Q.; Feng, Q. Multi-arm robot task planning for fruit harvesting using multi-agent reinforcement learning. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2023 Oct 1-5; Detroit, USA. IEEE; 2023. pp. 4176-83. DOI

108. Wete, E.; Greenyer, J.; Kudenko, D.; Nejdl, W. Multi-robot motion and task planning in automotive production using controller-based safe reinforcement learning. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems; Auckland, New Zealand. 2024. pp. 1928-37. Available from: http://jgreen.de/wp-content/documents/2024/AAMAS_24_Multi-Robot-Motion-and-Task-Planning-in-Automotive-Production-Using-Controller-based-Safe-Reinforcement-Learning.pdf. (accessed 2025-01-21).

109. Liu, Y.; Palmieri, L.; Georgievski, I.; Aiello, M. Human-flow-aware long-term mobile robot task planning based on hierarchical reinforcement learning. *IEEE. Robot. Autom. Lett.* **2023**, *8*, 4068-75. DOI

110. Li, X.; Yang, Y.; Wang, Q.; et al. A distributed multi-vehicle pursuit scheme: generative multi-adversarial reinforcement learning. *Intell. Robot.* **2023**, *3*, 436-52. DOI

111. Li, D.; Hou, Q.; Zhao, M.; Wu, Z. Reliable task planning of networked devices as a multi-objective problem using NSGA-II and reinforcement learning. *IEEE. Access.* **2022**, *10*, 6684-95. DOI

112. Zhang, J.; Ren, J.; Cui, Y.; Fu, D.; Cong, J. Multi-USV task planning method based on improved deep reinforcement learning. *IEEE. Internet. Things. J.* **2024**, *11*, 18549-67. DOI