小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0085
2026 年 4 月 第 4 期
Vol.47 No.4 2026

融合多模态感知的机器人抓取策略研究

禹鑫燚,何 威,欧林林

(浙江工业大学 信息工程学院,杭州 310023)
E-mail:linlinou@zjut.edu.com

摘 要: 物体抓取是机器人的基本技能,而在复杂场景中实现多样的操作抓取是一项具有挑战性的任务.针对复杂操作任务中机器人抓取系统面临的认知局限与动态场景适应性问题,本文提出了任务自适应的多模态感知融合框架.首先,结合视觉语言模型的图像理解和语义推理以及图像分割模型的检测识别,构建了多模态信息感知模型,实现任务场景的图文推理和物体识别.其次,融合语言提示和视觉提示提出了动态任务链分解机制,根据场景复杂程度实时调整任务操作步骤的分解并增强感知模型对图像物体的视觉理解.其次,针对机器人末端平行夹爪需适应不同场景的抓取任务问题,提出了一种视觉引导的抓取姿态优化网络,通过引入 2×2 网格策略进行抓取点预测以及编码器-解码器架构的姿态优化网络,联合优化姿态的几何精度与物理可行性.最后,为了快速适应不同场景下新工具或新物体抓取操作的零样本泛化任务,提出策略优化架构,综合考虑子任务完成、路径平滑性和时间效率,设计多维度奖励函数,使机器人能够适应动态环境并实时调整策略.通过设计复杂操作任务场景进行机器人的抓取实验,证实了所提的方法在不同应用场景的操作扩展性能,对于各种复杂抓取任务有着良好的泛化性和鲁棒性.

关 键 词: 多模态感知;机器人抓取;策略优化;抓取姿态优化
中图分类号: TP242 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0894-08

Research on Robot Grasping Strategy Integrating Multimodal Perception

YU Xinyi, HE Wei, OU Linlin
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Object grasping is a basic skill of the robot, while achieving versatile manipulation in complex scenarios remains a challenging task. This paper proposes a task-adaptive multimodal perception fusion framework to address cognitive limitations and dynamic scenario adaptability challenges in robotic grasping systems for complex manipulation tasks. First, a multi-modal information perception model is constructed by integrating the image comprehension and semantic reasoning capabilities of vision-language models with the detection and recognition functions of image segmentation models, thereby achieving textual-visual reasoning and object identification in task scenarios. Subsequently, a dynamic task chain decomposition mechanism is developed through the fusion of linguistic and visual prompts, which enables real-time adjustment of operational step decomposition according to scenario complexity while enhancing the visual comprehension of image objects by the perception model. To address the requirement for parallel grippers at robotic end-effectors to adapt to diverse grasping tasks across scenarios, a visually-guided grasping pose optimization network is proposed. This network incorporates a 2×2 grid strategy for grasp point prediction and an encoder-decoder architecture-based pose optimization network to jointly enhance geometric precision and physical feasibility of grasping postures. Finally, for rapid adaptation to zero-shot generalization tasks involving novel tools or objects in varying scenarios, a policy optimization architecture is established. This architecture comprehensively considers subtask completion, trajectory smoothness, and temporal efficiency through a multi-dimensional reward function design, enabling robotic systems to dynamically adjust strategies while adapting to environmental changes. Experimental evaluations conducted in complex manipulation scenarios demonstrate that the proposed framework exhibits scalable operational performance across diverse applications, achieving robust generalization capabilities and adaptability to intricate grasping tasks.

Keywords: multimodal perception; robotic grasp; policy optimization; grasp pose optimization

0 引 言

随着大规模语言模型(LLM)的进步,机器人技术取得了重大创新[1].LLM 在处理复杂任务方面具有强大的能力,包括感知人类意图、理解语义以及通过理解和生成自然语言指令来规划任务[2].深度学习、强化学习等人工智能技术的快速发展为机器人任务规划问题提供了可行的解决方案,取得了令人瞩目的研究成果[3].对于大多数简单的抓取任务,基

收稿日期:2025-02-26 收修改稿日期:2025-04-25 基金项目:浙江省自然科学基金白马湖实验室联合基金项目(LBMHD24F030002)资助;国家自然科学基金项目(62373329,62203392)资助;浙江省高层次人才特殊支持计划项目(2021R52004)资助. 作者简介:禹鑫燚,男,1979 年生,博士,副教授,研究方向为数字孪生、智能机器人;何 威,男,2000 年生,硕士研究生,研究方向为智能机器人;欧林林,女,1980 年生,博士,教授,研究方向为智能机器人.

于 LLM 的规划,机器人可以相对容易地完成[4]. 然而,当面临单靠夹爪无法完成抓取的任务时,该方法仍存在局限性[5]. 当人类或动物面临无法直接用手解决的问题时[6],他们可以选择合适的工具,例如使用开瓶器或剪刀打开快递箱. 重要的是,虽然人类可以根据现有知识选择合适的工具,但机器人不具备这种特性来有效地解决当前的问题[7,8]. 因此,机器人抓取系统需要具备根据场景和任务目标推断辅助工具的能力[9]. 或者,当没有合适的工具时,它们应该建议人类提供其他合适的工具. 在以前的研究中,当机械臂在直接抓取目标时遇到困难时,通常采用两种主要方法:提供或设计特定工具进行辅助,或修改夹持器机制以增强抓取能力.

在本文中,探讨了机器人如何针对此类场景提出工具选择解决方案,并探索如何适当地抓取工具以满足使用标准并辅助完成任务. 机器人在现实场景中选择工具进行辅助抓取,而智能代理应满足 3 个基本条件. 首先,它应该能够仅根据场景和指令信息进行学习并生成策略,而无需明确的指导. 其次,当没有最佳辅助工具时,它应该能够选择满足任务要求的其他工具. 最后,用于抓取的工具应遵守使用标准. 本文的贡献可以体现在以下几个方面:

1)构建图像视觉和指令文本多模态信息的联合感知模型,实现对场景图像的分割检测和指令文本的语义分析,预测目标物体和对应的辅助工具.

2)对于复杂任务中存在的长时序操作流程,结合物体的物理特征和场景的状态信息特征,提出了一种动态任务链分解的操作策略规划方法,融合任务和思维链的语言提示分类标签的图像提示构建多模态的提示方法,生成合适的子序列操作步骤.

3)基于 2×2 网格划分和抓取点预测的方法,结合生成工具的初始抓取姿态,并设计以编码器-解码器为主要架构的姿态优化网络,提取初始姿态的点云特征进行融合解码生成优化的姿态,从而实现工具的规范抓取,减少后续操作步骤的失误.

4)针对机器人使用工具辅助操作流程中的路径规划问题,设计了基于近端策略优化算法的反馈方法,可以有效处理连续动作空间并适应动态场景下的策略更新.

最后,本文在现实世界的机器人 Rokae XmateER7 Pro 上训练策略并部署整个抓取方法,通过设计多组不同的复杂实验场景和机器人操作任务,与其他方法进行对比实验以及消融实验,从而验证该方法在现实环境的有效性和泛化性.

1　相关工作

1.1　工具辅助抓取方法

先前的研究已经提出了几种使机器人学习使用工具的方法[10]. 当机械臂遇到目标超出其抓取范围的任务时,可以使用 L 形工具将物体拉回[11]. 这涉及通过无监督生成训练任务来学习稳健的技能[12],以及在模拟中使用不同的程序生成的任务来训练可转移技能以解决现实世界的顺序操作任务[13]. 此外,可以根据设计策略输出的任务信息来设计工具,从而快速创建适合任务的原型工具,并使用具有联合学习策略的强化学习框架来完成操作任务[14]. 在文献[15]中,提出了一种能够在杂乱环境中抓取和识别已知和新物体的机器人拾取系统.

该系统的特点是具有可伸缩机构的新型多功能夹持器,可在吸力和抓取之间快速切换. 在文献[16]中,提出了一种基于工具的参数化操作策略,用于处理同一表示空间内夹持器、物体和工具之间的交互. 然而在实践中,工具的选择会受到物体材质的影响. 例如在切水果时,塑料刀和金属刀的含义不同,后者具有优势. 因此机器人应该在场景中比较工具的材质和形状,评估物体的物理特性,并确定最佳抓取姿态[17]. 使用工具时,抓取工具必须满足任务和场景的要求,以符合使用标准. 不正确的使用可能导致任务失败甚至损坏场景[18]. 因此,需要对姿态生成和运动轨迹进行适当的规划[19].

1.2　基于任务的策略规划方法

在日常生活中,机器人通常需要执行涉及多个步骤和不同环境条件的长序列操作. 这需要对物体的位置进行准确的实时感知,以及在复杂环境中进行路径规划和避障[20]. 通过提供适当的任务提示,可以将复杂的操作任务或长序列空间任务分解为更简单的子任务,从而提高整体任务的效率和成功率[21,22]. 文献[23]提出 LLM 根据给定数据集中的可供性生成可行的操作任务,并为每个分解后的子任务建立奖励函数. 这种方法涉及训练和学习策略以及互连以处理长序列任务. 然而,任务分解需要高效且实用,不相关的步骤和过于复杂的子任务只会增加任务完成时间并降低系统的整体效率.

在文献[24]中提出了一种利用目标条件模仿学习并训练编码器从观察中提取一系列潜在子目标的方法,探索了序列预测方法在长序列规划和多任务决策中的应用. 然而,这种方法并没有解决数据集中未包含的新任务的问题,需要重新规划有效的策略.

1.3　基于强化学习的反馈方法

具身智能系统需要根据给定的指令进行全面的任务规划. 然而,在执行过程中,环境影响、物体遮挡、物体本身的特性等因素会对复杂任务的完成构成挑战.

强化学习方法作为反馈机制,可用于学习执行这些子任务的最佳策略[25]. 当应用于机器人策略时,大型模型不仅可以处理多模态动作分布,而且还表现出稳定的训练过程[26]. 因此,强化学习和大型模型的结合对于学习机器人运动策略具有重要的研究价值[27]. 在文献[28]中,各种机器人技能都是通过生成模拟自动学习的,利用基础模型和生成模型自动创建不同的任务、场景和训练监督,而对于抓取任务,采用 SAC 算法[29]. 文献[30]建议利用大型语言模型来定义奖励函数,以优化和完成各种机器人任务. 奖励充当与 LLM 的中介接口,有效地弥合了高级语言指令和低级机器人动作之间的差距.

2　方　法

2.1　系统概述

在复杂操作任务下,机器人可能无法直接对任务目标进行抓取,例如物体处在机器人的工作范围之外、对于目标物体的位置无法生成合适的抓取姿态以及物体形状不适合抓取等情况的存在,为了解决这些问题,需要机器人在任务中学会使

用一些工具进行辅助完成. 本文提出了一种复杂任务中机器人智能操作的方法,整体结构如图 1 所示.

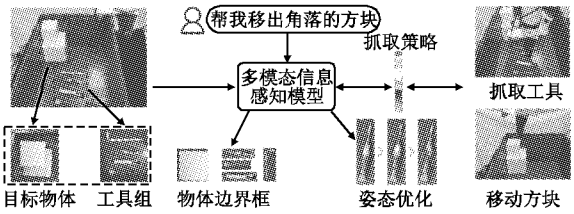


图 1 系统结构图

Fig. 1 System structure diagram

当输入自然语言指令和任务场景图像时,图像通过 SAM 模型对场景中的物体进行分割检测,输出对应的物体边界框和物体掩码图像,其中采用 ViT-H 作为预训练模型,用于图像分割任务. 将文本指令和处理后的图像传递给 GPT-4V,结合文本和图像的特征信息并对齐两者之间的语义空间,预测出对应的目标物体和辅助工具. 同时,经过提示的任务链分解,将指令分解为多个子任务操作序列,从而完成输入多模态信息的感知和处理,为后续策略规划提供信息基础.

2.2 多模态信息感知模型

现有的多模态模型如 GPT-4V,虽然能够通过图像与文本的联合输入进行语义推理,但其在像素级图像定位和几何特征解析方面仍存在不足,例如在复杂场景中,GPT-4V 可能无法精准区分堆叠物体的边界或与文本指令关联图像中具体的物体. 这种局限性直接影响了机器人对环境的感知精度以及后续的任务规划和决策生成环节.

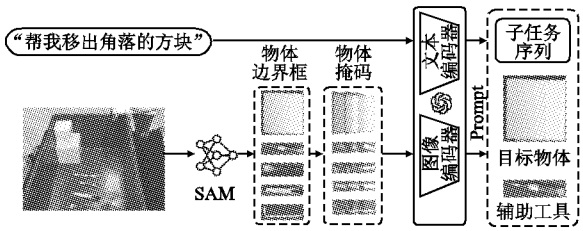


图 2 多模态信息感知模型结构图

Fig. 2 Structural diagram of the multimodal information perception model

SAM(Segment Anything Model)模型是一种新型的图像分割模型,能够根据自然语言指令和图像输入以分割掩码的形式准确地输出目标对象的精确位置和形状. 因此多模态模型可以结合 SAM 的优势去弥补图像定位的不足,形成具有多模态信息联合感知能力的视觉语言模型框架,如图 2 所示. 当任务场景图像时,图像通过 SAM 模型对场景中的物体进行实例分割,输出对应的物体边界框和掩码图像,其中采用 ViT-H 作为预训练模型具有零样本泛化能力,用于图像分割任务时无需针对特定物体进行微调. 分割结果进一步编码为几何特征向量,包括物体中心坐标,为后续抓取姿态优化提供关键几何约束. 将文本指令和处理后的图像传递给 GPT-4V,通过跨模态注意力机制对齐视觉特征与语言语义,预测出对应的目标物体和辅助工具. 最后通过任务链分解,将指令分解为多个子任务操作序列,从而完成输入多模态信息的感知和处理,为后续策略规划提供信息基础.

2.3 融合视觉语言的提示构建

对于大型语言模型,精心设计的提示工程有助于其更好地理解和分析任务,将复杂的语言指令转换为可操作的子任务序列. 在前人的工作中,提示工程一直是具身智能研究的重要焦点. 在“代码即策略”[31]中,代码功能被创新性地融入到提示工程中,并提出了将语言转换为策略代码的模型. 而在文献[32]中,加入了物体位置信息和交互关系,以辅助模型进行机器人的避障路径规划. 在此基础上,本文将物体状态和物理属性融入到语言提示工程中,如图 3 所示. 物体状态是指场景中目标物体的上下文关系,比如目标物体位于箱子内部的一角. 物理属性包括基本尺寸、重量、材质和基本动作原语. 描述物体状态有助于模型了解当前场景中物体的具体情况,指导模型判断工具的使用情况并选择合适的工具. 材质属性的引入旨在确保所选的用于工具辅助抓取的工具不会对任务目标物体造成损坏.

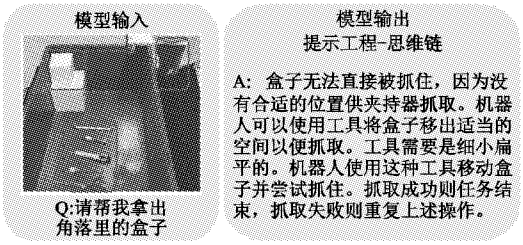


图 3 语言提示

Fig. 3 Language prompt engineering

此外,为了提高模型在复杂场景任务中的推理表现,在提示工程中引入了以任务和思维链(CoT)[33]为基础的提示构建方法. 在输出最终结果之前,会给出中间的逐步推理步骤,包括对象状态分析、工具选择、工具使用和后续操作.

为了确保分解的子任务操作序列适应动态场景的变化或前述任务失败造成的场景变化,在提示中根据场景复杂度动态调整任务分解,避免初始分解对场景的不适配性以及步骤冗余或不足. 基于 Doersch 等人[34]提出的根据上下文推理的自监督学习方法,在图像中采用拼图和标注的方式来构造辅助任务的想法,提出基于视觉的提示方法,如图 4 所示. 首先,通过 SAM 模型对场景图像进行分割,对图像中分割出的所有物体进行标注来形成视觉提示. 为场景中每个物体分配一个唯一编号,并根据模型提供图像中物体中心所在的坐标点进行注释编号标记,为后续推理提供更完善的信息.

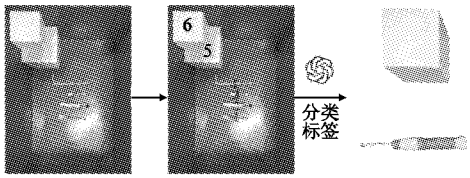


图 4 图像提示

Fig. 4 Vision prompt engineering

通过将复杂问题分解为多步骤子问题并按顺序解决,大型语言模型可以更好地理解任务场景并制定合理的策略计划. 语言模型生成子任务序列,检测潜在失败点,如未成功抓

取工具,更新任务序列.

2.4　基于 2×2 网格策略的抓取点预测

通过输入的物体图像边界框,2×2 网格策略将物体的抓取点选择转变为在 2×2 网格内选择最佳区域,增强了整体方法处理低分辨率图像的能力结合更广泛的上下文信息,在较低的像素密度下使抓取位置的选择过程具有鲁棒性. 网格将物体图像边界框划分为 2×2 个基本网格块,每个网格块都根据安全性、稳定性和可抓取性等标准进行评估. 结合 SAM 模型和 GPT-4V 对目标物体的预测分割,在网格内输出首选的抓取位置,从而指引后续的物体图像分割和姿态生成. 不同于依赖单个最佳抓握姿势选择的传统方法. 首先根据初始抓取姿态与首选位置的接近程度评估多个候选抓取姿态,再从中选择得分最高的姿态. 通过 2×2 网格策略相结合,确定最佳抓握区域,确保所选抓取姿态的稳定性和最佳性能,从而提高整体抓取方法的性能和成功率.

2.5　抓取姿态优化

在完成工具抓取步骤的过程中,机器人经常会遇到抓取位置不合适的问题,从而导致抓取失败,因此生成的抓取姿态对于工具能否抓取成功并辅助完成任务至关重要. 针对这种情况,本文将上一步提供的抓取预测点作为监督信号,提出了抓取姿态改进的流程,网络架构如图 5 所示.

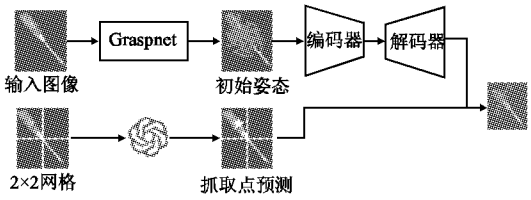


图 5　姿态优化网络结构图

Fig. 5　Structure diagram of the pose optimization network

当给出场景中辅助工具对应的 RGB-D 图像输入,由 Graspnet 模型[35]生成初始的一组候选姿态. 根据编码器与解码器联合的模块结构,并结合 GPT-4V 对网格化图像的抓取点预测,输出优化后工具的六自由度抓取姿态,从而使姿态符合工具使用规范. 在该网络中,编码器负责从高维输入数据中提取复杂特征. 编码器对点云数据和初始姿态进行预处理,并将这些信息转换为一组高级特征,为后续的抓取姿态改进过程提供基础.

解码器用于将编码器提取的特征向量转换为具体的抓取姿态参数. 它接收点云和初始姿态作为输入,通过编码器提取特征,对特征进行融合并解码以生成改进的抓取姿态. 同时,初始抓取姿态由 GraspNet[35]生成. 该网络由编码器和解码器组成. 在编码器中,数据预处理模块通过体素网格过滤点云数据 P_i,并将其映射到稀疏的三维网络中. 经过滤的点云表示为 P_i'. 对于点云 P_i'中的每一个点 p_{ii}',基于多层感知机的特征提取模块通过对所有点特征进行最大池化操作来提取全局特征 F_{P_i},函数表示为:

$$F_{P_i} = \underset{p_{ii}' \in P_i'}{max} MLP(p_{ii}') \tag{1}$$

初始抓取姿态 G 包括夹爪的空间位置 $p \in \mathbb{R}^3$和旋转 $r \in \mathbb{R}^3$,姿态特征 F_G 是通过编码函数 Φ_G 获得的. 编码函数 Φ_G 由多个全连接层组成,这些全连接层将机器人的位置和旋转转换为深度特征表示:

$$F_G = \Phi_G(p, r) \tag{2}$$

综合特征 F 是通过融合全局特征 F_{P_i}和姿态特征 F_G 获得的,它为优化抓取姿态提供了足够的上下文信息. 映射过程结合了编码特征 F 中的信息,并构建了一个能够表达抓取姿态的高级特征表示:

$$H_i = \sigma(W_i H_{i-1} + b_i) \tag{3}$$

其中,H_i 是前一层的特征表示,σ 是非线性激活函数. W_i 和 b_i 分别为权重矩阵和偏置向量. 为了重建夹爪的位置和方向,解码器通过非线性映射生成一个预测的抓取姿态 $\hat{G} = (\hat{p}, \hat{r})$,位置 $\hat{p}$ 和方向 $\hat{r}$ 的函数输出分别为:

$$\hat{p} = tanh(W_p H_l + b_p) \tag{4}$$

$$\hat{r} = N(tanh(W_r H_l + b_r)) \tag{5}$$

其中,W_p 和 W_r 是权重矩阵,b_p 和 b_r 是偏置项,H_l 是解码器最后一层隐藏层的输出. N 是归一化函数,用于确保四元数的单位性质. 为了准确量化模型预测的抓取姿态与实际抓取姿态之间的差异,损失函数设计如下:

$$L(G, \hat{G}) = \alpha \| p_t - \hat{p} \|_2^2 + \beta \left(1 - \left(\frac{< r_t, \hat{r} >}{\| r_t \| \, \| \hat{r} \|}\right)^2\right) \tag{6}$$

位置损失通过计算预测位置 $\hat{p}$ 和真实位置 p_t 之间的欧几里得距离的平方来衡量,这代表了模型在空间定位精度方面的性能,损失函数表示为:

$$L_{pos} = \| p_t - \hat{p} \|_2^2 \tag{7}$$

方向损失通过计算四元数的余弦相似度的平方来衡量,这能有效反映预测旋转与实际旋转之间的一致性,函数为:

$$L_{orient} = 1 - \left(\frac{< r_{true}, \hat{r} >}{\| r_{true} \| \, \| \hat{r} \|}\right)^2 \tag{8}$$

该损失项确保预测的抓取姿态方向尽可能接近实际姿态,这对保证抓取成功率至关重要. 权重参数 α 和 β 用于平衡位置误差和方向误差. 选择合适的权重参数是优化损失函数的关键. 在训练过程中,会根据每个 batch 的性能反馈调整 α 和 β 的值,实现自适应优化.

2.6　操作策略优化

在现实世界中,当没有提供适合任务的工具时,人类会根据当前情况选择一个合适的工具来协助完成任务. 然而,机器人并不具备这种特性. 因此,本文设计了一种基于近端策略优化(PPO)[36]的联合训练策略,PPO 是一种基于策略的强化学习算法,能够通过与环境的交互学习最优的行为策略,能够适应动态环境并实时调整策略以应对环境变化,可以很好地解决上述问题.

首先,对任务对应的环境进行建模,将机器人的状态、工具的状态以及目标物体的状态等信息编码为状态向量,表示为:

$$S_t = [S_{robot}(t), S_{tool}(t), S_{object}(t)] \tag{9}$$

其中,S_{robot}为机器人的状态,包括机械臂的六轴关节角、末端执行器的位姿;S_{tool}为工具的位姿,以及是否被抓取;S_{object}为任务对应的目标物体在场景中的位置和是否被移动. 其次,定义机器人的动作空间,包括机械臂的运动、工具的使用以及对目标物体的操作,表示为:

$$a_t = [a_{robot}(t), a_{tool}(t), a_{object}(t)] \tag{10}$$

其中,$a_{robot}(t) = [\Delta\theta_1(t), \Delta\theta_2(t), \cdots, \Delta\theta_7(t)]^T$ 为机械臂的

实时运动，$\Delta\theta_i(t)$表示第 i 关节在时间 t 处的角度变化，且每个关节都有对应的动作范围，$a_{robot}\in\mathbb{R}^6$. $a_{tool}(t)$的取值范围为{0,1}，目标物体操作 $a_{object}(t)$ 包括移动和旋转，为六维向量.

当机器人将工具移动到目标物体对应的操作位置并匹配目标姿态时，给予正奖励，设定奖励函数为：

$$R_{complete}=\begin{cases}1 & \|p_{current}-p_{target}\|\leqslant\epsilon_p,\Delta\theta\leqslant\epsilon_\theta\\ 0 & otherwise\end{cases}\tag{11}$$

其中，ϵ_p 为位置误差阈值，设置为 0.01m，$\Delta\theta$ 为姿误差，以计算四元数之间的夹角 $\Delta\theta=2\arccos(|q_{current}\cdot q_{target}|)$，$\epsilon_\theta$ 为姿态误差阈值，设置为 0.1 rad. 为了保证机械臂在操作过程中运动轨迹的平滑性，通过加速度标准差衡量路径的抖动程度，函数表示为：

$$R_{smooth}=-k_{smooth}\cdot\sigma_a\tag{12}$$

其中，k_{smooth}为平滑性权重，设置为 0.1，σ_a 为过去 N 个时间步的加速度标准差：

$$\sigma_a=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i-\bar{a})^2},\bar{a}=\frac{1}{N}\sum_{i=1}^{N}a_i\tag{13}$$

为了鼓励机器人在最短时间内完成任务，避免耗时过长，引入时间惩罚，表示为：

$$R_{time}=-k_{time}\cdot\frac{t}{T_{max}}\tag{14}$$

其中，k_{time}为时间惩罚权重，设置为 0.1，t 为任务开始到当前的时间，T_{max}为任务最大允许时间，设置为 30s. 综合上述奖励和惩罚项，总奖励函数为：

$$R=R_{complete}+R_{smooth}+R_{time}\tag{15}$$

PPO 算法以最大化累积奖励直接优化策略参数. 这种方法在实现高效更新的同时，确保了算法的稳定性. 网络参数的更新方式如下：

$$L^{CLIP}(\theta)=\hat{E}_t[min(r_t(\theta)\hat{A}_t,clip(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t)]\tag{16}$$

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\tag{17}$$

其中，ϵ 表示截断超参数，设置为 0.2，用于限制策略更新的幅度. $r_t(\theta)$为概率比率，用于衡量新旧策略的差异. $clip$ 函数是一个截断函数，它将比率 r_t 限制在范围$[1-\epsilon,1+\epsilon]$内，以确保收敛，L^{CLIP}使用 min 函数选择较小的值作为目标的下界. 此外，基于广义优势估计(GAE)优势函数，能有效降低估计的方差并提高学习的稳定性. 参数 λ 设置为 0.95，该函数的形式为：

$$\hat{A}_t^{GAE(\gamma,\lambda)}=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1}\tag{18}$$

$$\delta_t=r_t+\gamma V(S_{t+1})-V(S_t)\tag{19}$$

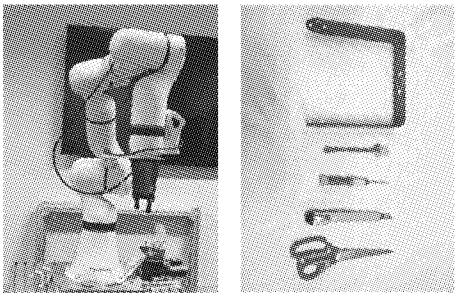
3 实 验

3.1 模型训练

姿态网络在 PyTorch 框架中实现，并使用 NVIDIA RTX 4090 Ti GPU 上的 AdamW 优化器进行训练. 姿态网络包括姿态生成和姿态优化. 这两个网络使用不同的超参数和数据集进行相对训练. 姿态生成网络在 grassnet-1billion 数据集上进行训练，以物体局部点云为输入，输出完整的物体点云和抓取姿态. 本文将 batch size 设置为 8，训练网络 300 个 epoch. 根据 GraspNet 生成的抓取姿态，将姿态的随机选择改为根据物体重心位置进行选择，以保证抓取过程中物体不会掉落. 姿态优化网络的训练数据集是根据实验过程中收集的抓取状态创建的，包括 grassnet-1billion 数据集中的大多数物体. 该网络以点云和抓取姿态为输入，输出改进的点云和抓取姿态.

3.2 实验准备

为了验证抓取方法的有效性，本文进行了真实环境的实验以评估抓取技术、物体材料对抓取的影响以及现实世界的抓取场景. 如图 6(a)所示，实验使用了珞石机器人的 ER7 Pro 机械臂和钧舵的 RG50 夹持器，使用 Realsense D435 相机捕获场景的 RGB-D 信息.



(a)机械臂 (b)测试工具：框型工具、刷子、小螺丝刀、美工刀、剪刀(从上到下)

图 6 实验环境

Fig. 6 Experimental environment

在实验中，本文设计了几个任务来展示在工具选择和姿态优化方面的优势，以完成任务. 在场景中，本文设计了桌面抓取场景，其中每个任务涉及的物体和工具都是随机生成的.

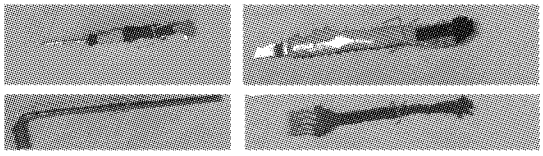
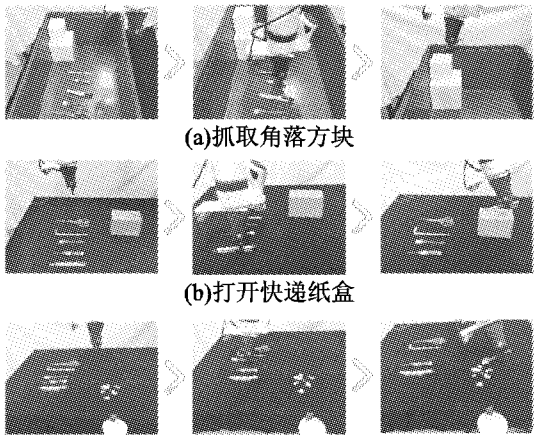


图 7 工具抓取示意图

Fig. 7 Tool grasp pose diagram



(a)抓取角落方块

(b)打开快递纸盒

(c)清理桌面纸屑

图 8 真实环境抓取效果图

Fig. 8 Grasp renderings in the real environment

生成的工具在固定的起始位置初始化. 工具抓取的可视化实

验效果如图 7 所示,真实环境中的抓取实验效果如图 8 所示. 这些任务的简要描述如下:

a)移动盒子进行抓取:利用工具移除位于盒状容器角落的无法直接抓取的物体. 将刮刀插入物体和容器壁之间的缝隙中,操纵刮刀将物体移出角落,从而便于正确抓取.

b)打开快递箱:选择工具打开用胶带密封的包裹箱. 将刀或锋利的工具放在包裹箱上方,确保有足够的高度来切开胶带. 切开胶带后,尝试从盒子中取出里面的物体.

c)清理桌面:使用工具清理桌面上的杂物. 将收集容器放在桌子边缘,并利用末端执行器夹持器抓取刷子将碎屑扫入容器中.

3.3　评　估

为了评估抓取方法在抓取任务中的有效性,本文对工具选择网络和姿态优化网络进行了定性分析实验. 为了验证抓取方法在抓取任务中的优越性能以及在不同场景的方法泛化性能,本文设计了与其他抓取方法在不同场景下的比较实验. 最后,为了评估抓取方法中每个模块对抓取性能的影响,本文设计了消融实验来验证每个模块的必要性.

首先,采用不同的视觉语言模型与本文方法进行对比,比较提示工程对于模型预测辅助工具以及生成操作步骤的性能. 引入平均精度,即综合考虑检测的准确性和完整性,计算平均精度,评估模型在图像中检测和定位目标对象的能力. 引入视觉问答准确率(VQA)评估模型根据图像内容以及语言指令回答问题的能力,即模型生成的步骤与标准步骤一致的比例. 最后,计算模型对于步骤生成的推理时间来评估计算效率. 在实验过程中,随机给出 5 种不同工具以及场景中存在的文本任务进行 30 次实验,实验数据如表 1 所示.

表 1　真实环境中不同模型对比结果

Table 1　Compare the results of different models in real environment

方法	mAP	VQA	平均推理时间(s)
BLIP[37]	0.756	0.833	8.8
CLIP[38]	0.724	0.833	9.4
Our	**0.891**	**0.967**	**7.2**

在复杂场景中,模型的性能受到目标检测能力、视觉问答能力以及计算效率的影响,而从表中数据可以看出,本文方法在复杂场景下能够更准确地检测和定位目标对象,模型基于思维链的提示能够更准确地理解图像内容,并生成正确的操作步骤,从而进一步提高推理效率,减少计算时间.

其次,为了验证优化抓取姿态的可行性,本文根据生成的姿态进行工具抓取,并对抓取结果进行定量分析. 使用两个指标来评估抓取姿态:抓取的成功率和抓取后工具是否处于正常操作状态. 通过将工具分为功能区和抓握区,选定抓握区中心 ±10% 抓握区总长度作为合理抓取区间,生成的抓取姿态在区间内则认定为成功,以姿态符合率作为评估指标. 随机选取 3 种工具,通过对比其他抓取姿态的生成方法进行 30 次抓取验证实验,以体现本文方法的性能.

由表 2 数据可以看出,改进的姿态生成方法在保证较高抓取成功率的前提下,能够更好地结合物体的几何特征和场景信息,生成符合工具操作规范的抓取姿态,从而显著提高了工具被抓取后使用的可行性,在实际应用中更具实用性和可靠性.

表 2　工具抓取不同方法对比结果

Table 2　Comparison of different methods of tool grasping results

方法	抓取成功率(%)			姿态符合率(%)		
	刀具	勺子	刷子	刀具	勺子	刷子
Mousavian[39]	76.7	73.3	70.0	70.0	76.7	73.3
GraspNet[35]	83.3	86.7	87.7	73.3	70.0	76.7
Our	**93.3**	**93.3**	**90.0**	**90.0**	**93.3**	**93.3**

为了验证机器人抓取系统在真实的杂乱环境下针对目标物体抓取任务中的性能. 如图 9 所示,给定初始语言指令“给我一个 L 型工具”,并设置 3 种实验场景,分别摆放 3、6、10 个不同的常见物体,其中包含任务对应的目标物体,机器人初始化与上节相同. 每个场景均进行 30 次实验,并使用 3 个指标与 VLG[40]方法进行对比衡量:

1)任务成功率:定义系统在一次实验抓取任务中,尝试抓取的次数少于物体对应的数量并成功抓取目标物体,则该任务视为成功. 计算任务成功次数占总次数的比例.

2)平均抓取次数:每个任务的平均抓取动作数.

3)动作耗时:任务成功的平均时间. 失败的任务会导致抓取时间无意义的延长,因此只计算任务完成的平均耗时.

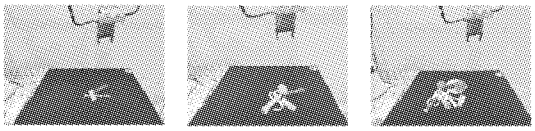


图 9　堆叠环境工具抓取实验

Fig. 9　Stacking environment tool grasping experiment

堆叠场景的工具抓取实验结果如表 3 所示,可以看出,面向物体较少的堆叠场景,可以实现较高的成功率,并且以较短的抓取时间和抓取次数完成任务,证明了方法具有良好的性能以及较好的抓取稳定性. 而当物体数量增多时,抓取成率、抓取次数以及抓取时间均表现变差,这是由于在真实场景中,抓取方法会受到物体的表面摩擦、质量属性,夹爪的夹持力度以及相机获取图像中噪声的影响.

表 3　堆叠场景抓取实验结果

Table 3　Experimental results of stacking environment

物体数量	抓取成功率(%)	抓取次数	动作耗时(s)
	VLG|**Our**	VLG|**Our**	VLG|**Our**
3	86.7|**93.3**	1.39|**1.14**	11.18|**9.23**
6	83.3|**86.7**	2.13|**1.56**	25.34|**14.76**
10	66.7|**76.7**	3.62|**2.39**	61.75|**31.33**

为了验证机器人抓取系统在混合场景下实现抓取的有效性,设计了堆叠场景抓取任务和工具辅助操作任务相结合的实验场景,如图 10 所示. 将目标工具随机放置在物体数量分别为 3、6、10 的杂乱场景中,机器人抓取系统通过给定的自然语言指令“获取场景边缘的梨子”以及相机获得的场景图像,确定目标物体和对应的辅助工具,由机器人操作子系统抓从杂乱场景抓取出辅助工具并完成操作任务. 进行 30 次综合实

验,通过计算不同物体数量下工具抓取的成功率、任务的总成功率、工具完成抓取的平均时间以及成功完成任务的平均时间评估方法的性能.其中,工具抓取的成功率为 30 次实验中成功抓取正确工具的次数,总成功率为成功完成整个任务的次数占工具抓取成功次数的比例,平均成功完成时间为完成整个任务的平均耗时.

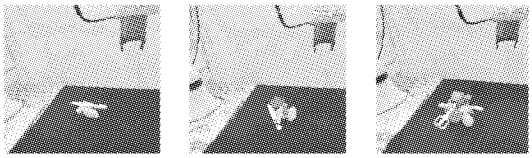


图 10 混合环境抓取实验

Fig. 10 Mixed scene grasping experiment

实验结果如表 4 所示,与杂乱场景的抓取情况相同,物体数量对于实验结果有着很大的影响,尤其是工具的抓取成功率.但是对比表中工具抓取的平均消耗时间,物体数量为 3 时,机器人使用工具的平均操作耗时为 24.6s;物体数量为 6 时,工具操作耗时为 23.7s;物体数量为 10 时,工具操作耗时为 24.5s.可以得出混合场景下,工具操作耗时受到物体杂乱程度的影响较小,在完成杂乱堆叠物体中的工具抓取后,以良好的鲁棒性完成后续的操作任务.

表 4 混合场景抓取实验结果

Table 4 Experimental results of mixed environment

物体数量	工具抓取成功率(%)	总成功率(%)	工具抓取耗时(s)	任务完成耗时(s)
3	93.3	92.9	10.1	34.7
6	86.7	88.5	14.5	38.2
10	73.3	81.8	29.8	54.3

最后,为了验证各个模块对系统整体性能的影响,通过消融实验来展现系统各部分对整体方法的影响,以图 8(a)为例,提供 5 种工具并进行 30 次实验,实验结果如表 5 所示.与完整方法相比,移除多模态感知模型将会严重影响工具识别和预测的成功率从而导致后续的任务失败;移除提示工程导致工具识别成功率下降并且任务完成的平均耗时增加了 3.8s,这是由于增加了模型对策略规划的推理难度,但未影响工具的抓取成功概率.当移除姿态优化网络时,并未影响对工具的识别和任务的成功完成,但是工具成功抓取的概率显著下降.而策略优化环节的移除会影响工具的操作路径导致任务成功率显著降低.

表 5 消融实验

Table 5 Ablation experiment

	工具成功识别概率	工具姿态合理概率	任务成功概率	任务完成耗时(s)
完整网络	**28/30**	**25/28**	**23/25**	**16.7**
移除感知模型	N	7/30	4/7	53.2
移除提示工程	14/30	11/14	7/11	20.5
移除姿态优化	27/30	12/27	9/12	15.3
移除策略优化	28/30	26/28	6/26	23.2

4 结束语

本文提出一种面向复杂操作场景的机器人抓取系统创新框架,通过多模态感知融合与动态优化机制突破传统方法的认知与适应性局限.基于视觉语言模型与图像分割的协同推理,构建多模态信息感知模型,实现对场景语义与物体属性的精准解析;提出的动态任务链分解机制通过语言-视觉联合提示调节任务粒度,结合场景复杂度实时生成可执行操作序列.针对抓取姿态优化问题,设计基于 2×2 网格策略的视觉引导网络,通过编码器-解码器架构联合几何特征与物理约束,实现六自由度抓取姿态的协同优化.进一步构建多维度强化学习策略架构,集成任务完成度、运动平滑性及时间效率的复合奖励机制,在动态环境中实现零样本的抓取任务泛化.实验表明,该方法在工具辅助操作、高遮挡杂乱场景中均有着良好的抓取性能,为机器人智能操作提供了可扩展的理论范式与技术基础.

References:

[1] Pateria S, Subagdja B, Tan A, et al. Hierarchical reinforcement learning: a comprehensive survey [J]. ACM Computing Surveys, 2021, 54(5): 1-35.

[2] Naveed H, Khan A U, Qiu S, et al. A comprehensive overview of large language models [J]. arXiv preprint arXiv: 2307.06435, 2023.

[3] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: a survey [J]. Science China Information Sciences, 2025, 68(2): 24824-24837.

[4] Hu Z, Ding Y, Wu R, et al. Deep learning applications in games: a survey from a data perspective [J]. Applied Intelligence, 2023, 53(24): 31129-31164.

[5] Chen D, Chen J, Fang C, et al. Complex visual question answering based on uniform form and content [J]. Applied Intelligence, 2024, 54(6): 4602-4620.

[6] Griffin D R. Animal minds: beyond cognition to consciousness [M]. University of Chicago Press, 2001.

[7] Baber C. Cognition and tool use: forms of engagement in human and animal use of tools [M]. CRC Press, 2003.

[8] Chandrasegaran S K, Ramani K, Sriram R D, et al. The evolution, challenges, and future of knowledge representation in product design systems [J]. Computer-Aided Design, 2013, 45(2): 204-228.

[9] Zhu Y, Zhao Y, Chun Zhu S. Understanding tools: task-oriented object modeling, learning and recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2855-2864.

[10] Zhu Z, Hu H. Robot learning from demonstration in robotic assembly: a survey [J]. Robotics, 2018, 7(2): 17-42.

[11] Brown S, Sammut C. Tool use learning in robots [C]//AAAI Fall Symposium: Advances in Cognitive Systems, 2011: 7842-7849.

[12] Tee K P, Li J, Chen L T P, et al. Towards emergence of tool use in robots: automatic tool recognition and use without prior tool learning [C]//IEEE International Conference on Robotics and Automation (ICRA), 2018: 6439-6446.

[13] Wang Y, Le H, Gotmare A D, et al. Codet5 +: open code large lan-

guage models for code understanding and generation[J]. arXiv preprint arXiv:2305.07922,2023.

[14] Liu Z,Tian S,Guo M,et al. Learning to design and use tools for robotic manipulation[J]. arXiv preprint arXiv:2311.00754,2023.

[15] Zeng A,Song S,Yu K T,et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching[J]. The International Journal of Robotics Research,2022,41(7):690-705.

[16] Noguchi Y,Matsushima T,Matsuo Y,et al. Tool as embodiment for recursive manipulation[J]. arXiv preprint arXiv:2112.00359,2021.

[17] Guo D,Xiang Y,Zhao S,et al. PhyGrasp:generalizing robotic grasping with physics-informed large multimodal models[J]. arXiv preprint arXiv:2402.16836,2024.

[18] Yu X,Huang R,Zhao C,et al. Def-grasp:arobot grasping detection method for deformable objects without force sensor[J]. Neural Processing Letters,2023,55(8):11739-11756.

[19] Yin H,Varava A,Kragic D. Modeling,learning,perception,and control methods for deformable object manipulation[J]. Science Robotics,2021,6(54):8803-8817.

[20] Li K,Chen J,Yu D,et al. Deep reinforcement learning-based obstacle avoidance for robot movement in warehouse environments[C]//IEEE 6th International Conference on Civil Aviation Safety and Information Technology(ICCASIT),2024:342-348.

[21] Zhang W,Cheng H,Hao L,et al. An obstacle avoidance algorithm for robot manipulators based on decision-making force[J]. Robotics and Computer-Integrated Manufacturing,2021,71:102114,doi:10.1016/j.rcim.2020.102114.

[22] Li S,Zhang S,Fu Y,et al. Task-based obstacle avoidance for uncertain targets based on semantic object matrix[J]. Control Engineering Practice,2020,105(11):104649,doi:10.1016/j.conengprac.2020.104649.

[23] Katara P,Xian Z,Fragkiadaki K. Gen2sim:scaling up robot learning in simulation with generative models[C]//IEEE International Conference on Robotics and Automation(ICRA),2024:6672-6679.

[24] Hong M,Kang M,Oh S. Diffused task-agnostic milestone planner[J]. Advances in Neural Information Processing Systems,2023,36(7):387-405.

[25] Shyalika C,Silva T,Karunananda A. Reinforcement learning in dynamic task scheduling:a review[J]. SN Computer Science,2020,1(6):306-313.

[26] Akalin N,Loutfi A. Reinforcement learning approaches in social robotics[J]. Sensors,2021,21(4):1292-1329.

[27] Zhang R,Lü Q,Li J,et al. A reinforcement learning method for human-robot collaboration in assembly tasks[J]. Robotics and Computer-Integrated Manufacturing,2022,73:102227,doi:10.1016/j.raim.2021.102227.

[28] Wang Y,Xian Z,Chen F,et al. Robogen:towards unleashing infinite data for automated robot learning via generative simulation[J]. arXiv preprint arXiv:2311.01455,2023.

[29] Hou Z,Zhang K,Wan Y,et al. Off-policy maximum entropy reinforcement learning:soft actor-critic with advantage weighted mixture policy(sac-awmp)[J]. arXiv preprint arXiv:2002.02829,2020.

[30] Yu W,Gileadi N,Fu C,et al. Language to rewards for robotic skill synthesis[J]. arXiv preprint arXiv:2306.08647,2023.

[31] Liang J,Huang W,Xia F,et al. Code as policies:language model programs for embodied control[C]//IEEE International Conference on Robotics and Automation(ICRA),2023:9493-9500.

[32] Huang W,Wang C,Zhang R,et al. Voxposer:composable 3D value maps for robotic manipulation with language models[J]. arXiv preprint arXiv:2307.05973,2023.

[33] Wei J,Wang X,Schuurmans D,et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems,2022,35(2):24824-24837.

[34] Doersch C,Gupta A,Efros A. A unsupervised visual representation learning by context prediction[C]//Proceedings of the IEEE International Conference on Computer Vision,2015:1422-1430.

[35] Fang H S,Wang C,Gou M,et al. Graspnet-1billion:a large-scale benchmark for general object grasping[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020:11444-11453.

[36] Gu Y,Cheng Y,Chen C L P,et al. Proximal policy optimization with policy feedback[J]. IEEE Transactions on Systems,Man,and Cybernetics:Systems,2021,52(7):4600-4610.

[37] Li J,Li D,Xiong C,et al. Blip:bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning,PMLR,2022:12888-12900.

[38] Radford A,Kim J W,Hallacy C,et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning,PMLR,2021:8748-8763.

[39] Mousavian A,Eppner C,Fox D. 6-dof graspnet:variational grasp generation for object manipulation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:2901-2910.

[40] Xu K,Zhao S,Zhou Z,et al. A joint modeling of vision-language-action for target-oriented grasping in clutter[J]. arXiv preprint arXiv:2302.12610,2023.