RoboDexVLM：基于视觉语言模型的灵巧机器人操作任务规划和运动控制

本文介绍了 RoboDexVLM，这是一个专为配备灵巧手的协作机械手量身定制的创新型机器人任务规划和抓取检测框架。以往的方法侧重于简化且功能有限的操作任务，往往忽略了以长视距方式抓取各种物体所带来的复杂性。相比之下，我们提出的框架利用灵巧手，能够抓取各种形状和大小的物体，并基于自然语言命令执行任务。该方法包含以下核心组件：首先，设计了一个强大的任务规划器，该规划器具有基于视觉语言模型 (VLM) 的任务级恢复机制，使系统能够解释和执行针对长序列任务的开放词汇命令。其次，基于机器人运动学和形式化方法，提出了一种语言引导的灵巧抓取感知算法，该算法专为零样本灵巧操作而设计，可处理多种物体和命令。全面的实验结果验证了 RoboDexVLM 在处理长视场场景和执行灵巧抓取方面的有效性、适应性和鲁棒性。这些结果凸显了该框架在复杂环境中的运行能力，展现了其在开放词汇灵巧操作方面的潜力。我们的开源项目页面位于 https://henryhcliu.github.io/robodexvlm。

一、引言

机器人操作已成为现代技术进步的基石，推动着制造业、医疗保健业和家庭自动化的发展。通过连接感知、推理和物理交互，这些系统提高了生产力，确保在危险环境中安全运行，并解决了劳动力短缺等重大社会挑战。机器人操控流程通常包含四个核心阶段：通过传感器（例如 LiDAR、RGB-D 摄像头）进行环境感知、通过物体检测定位目标、通过抓取感知计算稳定接触点，以及通过运动规划执行无碰撞轨迹。对于复杂、长期的任务，例如整理杂乱的空间或准备餐食，系统必须将抽象目标分解为可操作的序列，这需要对物体可供性和任务依赖性进行语义理解。

视觉感知领域的最新进展使得语言引导的物体检测和分割成为可能，这对于机器人操控至关重要。虽然像 YOLOv11 [1] 这样的传统检测器速度很快，但它们对标记数据集的依赖限制了其适应性。 Grounding DINO [2] 和 SAM [3] 等现代框架通过零样本泛化解决了这个问题：SAM 实现了与提示无关的分割，而 Grounding DINO 则将文本提示（例如“蓝色立方体”）与物体定位联系起来。像 LangSam [4] 这样的混合方法整合了这些模型，进一步实现了无需人工细化即可实时生成掩码，这对于动态操控任务至关重要。基于这些感知基础，近期的机器人操控研究专注于将分割后的物体转化为可执行的动作序列。

在机器人操作的任务规划方面，大型语言模型 (LLM) 的集成重新定义了机器人操作，使系统能够解读抽象指令并将其转化为可操作的序列 [5]。近期的研究利用视觉语言模型 (VLM) 将操作任务表示为约束满足问题 [6], [7]。例如，ReKep [8] 使用 VLM 将 3D 环境关键点映射到数值成本函数，通过最小化碰撞风险和不稳定性来优化抓取姿势。这种方法绕过了传统的基于启发式的规划器，从而能够高效地适应新物体。同样，OmniManip [9] 引入了一种以物体为中心的交互表示，将 VLM 衍生的语义推理与几何操作原语相结合。 OmniManip 通过构建用于规划和执行的双闭环系统，无需对 VLM 进行微调，即可使用并联夹爪实现跨多种任务的零样本泛化。RoboMamba [10] 进一步推进了这一范式，利用视觉语言-动作 (VLA) 模型，直接从语言提示预测目标物体的姿态和变换矩阵。CLIPort [11] 和 VoxPoser [12] 等互补方法展示了 VLM 如何分别合成空间可供性图或生成类似代码的动作脚本，从而引导机械手执行组装或厨房操作等复杂任务。这些创新共同凸显了向基于语言的操作的转变，其中语义理解和几何推理紧密结合。通过在 VLM 驱动的框架下统一感知、规划和执行，机器人系统能够灵活地解释模糊命令，以便在现实世界中部署 [13]。

然而，上述机器人操作方法仅仅使用了平行夹持器来简化具有高自由度 (DoF) 的具有挑战性的操作问题。相比之下，灵巧手以其多指设计和类似人类的关节 [14]，[15] 为特点，正在成为机器人在现实场景中自适应执行抓取操作的一种手段。与平行爪夹持器不同，平行爪夹持器擅长刚性、预先定义的抓取，但难以抓取精细或可变形的物体 [16]，灵巧手模拟人类操作的适应性 [17]。它们能够进行手部重新定位和接触丰富的交互，使其成为在混乱环境中使用工具或处理器具等任务不可或缺的工具。抓取感知方面的最新进展主要集中在平行夹持器上，产生了 GG-CNN [18]、Contact-GraspNet [19] 和 AnyGrasp [20] 等模型，它们根据点云或 RGB-D 数据预测对映抓取。相比之下，原生灵巧抓握感知需要精确的先验几何信息。诸如 D(R, O) 抓握 [21] 之类的方法统一了不同实施例中的机器人-物体交互表示，实现了跨平台抓握合成，但这严重依赖于精确的物体网格和关节扭矩约束。为了规避基于模型的规划的挑战，许多研究人员采用强化学习 (RL) 和模仿学习，并根据人类演示的动作捕捉数据进行训练 [22]，[23]。 DexCap [24] 等研究引入了可扩展的系统，用于通过 RL 策略收集高保真手部与物体交互数据，以实现目标条件下的灵巧抓取。尽管取得了这些进步，但仍存在一个关键的差距：当前的方法主要侧重于抓取姿势的生成，而忽略了灵巧操作与具身人工智能的任务级规划的整合 [25]，[26]。将灵巧操作与 VLM 驱动的任务规划器相结合代表着一个尚未开发的领域，其中零样本泛化和情境适应性可以释放机器人系统前所未有的多功能性。

最后，确保从故障中稳健恢复对于在现实世界中部署机器人操作系统至关重要，因为感知模糊性、环境不确定性和不完善的模型输出经常会导致错误 [20]、[27]。具体而言，AIC MLLM [28] 集成了测试时自适应功能，允许智能体根据所选场景中的实时分析动态调整其感知和规划模块。因此，自我评估机制有可能增强非结构化环境中的弹性 [9]，凸显了闭环自适应性对于可靠的长期机器人操作的重要性。尽管如此，在机器人操作中通过任务重新规划从故障中恢复仍然是一个有待持续探索的领域。

为了弥补上述研究空白，我们提出了 RoboDexVLM，这是一个用于灵巧操作的开放词汇任务规划和运动控制框架，如图 1 所示。通过将视觉数据与自然语言指令相结合，VLM 使机器人能够解读意图，根据给定的技能库推断任务层次结构，并动态调整计划。当与能够进行多种物体交互的灵巧手配合使用时，这种协同作用可以实现精确的、情境感知的操作，为机器人在非结构化的真实环境中无缝操作铺平道路。

贡献总结如下：

图 1. RoboDexVLM 概览。多模态指令（包含人类指令、可用技能列表、RGB-D 图像和相关记忆项）被传输到 VLM 进行任务规划。灵巧机器人接收到 VLM 发出的技能调用序列后，将执行这些技能直至任务完成。虚线表示操作失败后的恢复过程。

• 我们提出了 RoboDexVLM，这是一个新颖的框架，它将基于 VLM 的自动化任务规划流程与模块化技能库相结合，以实现长视域灵巧操作，从而有效地连接高级规划和低级运动学约束。

•该框架集成了可变语言学习器 (VLM)，支持基于基元的任务分解和执行，并支持开放词汇命令。强大的任务规划器能够动态解读用户意图，优化抓取姿势，并结合故障恢复机制，利用其反射能力实现长远的适应性。

•通过大量的实际实验，我们证明了其在复杂环境中的有效性，突出了其在灵巧抓取方面的稳定性、对新物体的适应性以及对不熟悉任务的弹性。

二、基于规范基元的语言操作

A. 基于操作基元的任务规划

RoboDexVLM 框架代表了机器人操作领域的重大进步，它通过创新的结构化技能库无缝衔接了高级任务规划和低级执行之间的鸿沟。该技能库是实现零样本操作能力的核心，机器人可以仅基于自然语言指令（例如，打开盒子并将更大的杨桃放入其中，如图 1 所示）执行未明确编程的任务。该技能库 S 的架构经过精心设计，旨在形式化操作基元，同时保持足够的灵活性，以允许在语言输入的指导下进行调整。在 S 中，每个技能单元 Fi 都有其所需的输入 Ii 来激活特定动作。

如图 2 所示，技能库 S 包含八项原子技能，涵盖了诸如检测、抓取、移动和放置物体等基本操作动作。这些原子技能 Fi 是构建 L 所指令的复杂操作任务的基石。每项技能都设计为在整体框架内独立运行，但又相互关联，从而确保任务的平稳过渡和高效执行。

利用世界知识和推理能力，可变学习模型 (VLM) 通过以下即时推理过程生成技能顺序和相应的所需功能输入：

其中，上下文生成器 K(·) 的输入元素为：包含思维链 (CoT) 推理模板的常量系统消息 S、记忆消息 Mτ 以及作为时间步 τ 的任务描述的人类消息 Lτ。我们将函数 T 表示为 VLM 的推理过程。VLM 代理的输出设计为三部分：CoT 推理文本 Rτ、待调用技能的原语顺序 Oτ，Oi ∈ S，以及相应的技能输入原语 I，其中 Ii ∈ set(Xi)。需要注意的是，Rτ 增强了 VLM 所设计技能顺序的透明度，并且对于小样本学习的记忆反射非常有益。

B. 技能执行的交互原语

RoboDexVLM 框架设计的一个关键方面是其针对每个技能的标准化输入输出接口 Fi(Xi)。这些接口促进了不同阶段之间的无缝交互，使 VLM 能够根据语言命令 Lτ 的特定要求动态地将它们链接在一起。因此，我们维护动态变量存储，其形式化如下：

E\_lang 表示图像分割的语言指导。RGB-D 图像的像素矩阵分别表示为 P\_RGB 和 P\_Depth。E\_lang 语义分割的二进制结果存储在 B\_image 中。对于物体抓取，G 包含必要的几何值，这将在第 IIIB 节中详细说明。与物体的最大接触力用 F\_max 表示。机器人运动（例如旋转和扭转）的几何矢量表示为 A = {d, θ, r} ∈ R3。此外，图 2 下方所示的技能函数 Fi 可以查询变量存储 D，以检索用于实时操作的更新数据。

图 2. RoboDexVLM 的工作流水线。该系统包含多个互补的模块，旨在构建闭环操作框架。任务管理器根据 VLM 生成的 O 来协调 Fi 的执行。技能的执行是通过系统核心建立的四项基础能力来实现的。

III. 灵巧操作的技能执行

结合技能 Oτ 的原始顺序，机器人系统按顺序执行灵巧操作，同时更新变量存储 D，直到完成初始语言指令 L 中的长周期任务。

A. 感知-动作范式

RoboDexVLM 框架的感知-动作范式旨在通过闭环执行系统实现精确且稳健的操作，如图 2 所示。该流水线集成了多种先进技术，以实现支持各种任务技能调用的基础能力。在每项基础能力中，RoboDexVLM 系统都遵循连贯感知动作范式。

首先，语言引导的图像分割模块通过将语言嵌入 E\_lang 与实时视觉输入 P\_RGB 相结合，生成语义级对象掩码。受 [4] 的启发，我们采用了两个互补的模型来生成语义掩码。首先，开放集对象检测器 Grounding DINO [2] 通过将 E\_lang（例如，红苹果）与从 P\_RGB 中提取的视觉特征对齐，执行零样本文本到边界框检测。计算边界框的区域提议得分 B\_i，并将超过阈值 τ\_d 的区域保留为 B∗。随后，SAM [29] 将 B∗ 细化为像素精度的掩码 B\_img，确保即使在杂乱或奇怪的场景中也能实现实例感知的分割。这种混合方法既提高了效率（通过由粗到精的处理减少了搜索空间），又增强了模糊指令下的场景理解鲁棒性。

其次，系统利用与分段掩码 B\_img 对齐的 P\_RGB 和 P\_Depth 来筛选目标物体，然后通过 AnyGrasp [20] 推断机器人末端执行器的最佳抓取姿势。对于每个候选抓取姿势假设 G\_j，几何-几何对齐通过余弦相似度计算成对对应分数：

其中 fθ(·) 通过学习到的网络 θ 将抓取姿势编码为特征向量。汇总各个假设之间的相似度形成对应矩阵 S\_pq ，其中 N 为候选数，候选 j 的原始置信度通过逐行求和得出：Cj = ∑\_k=1 S\_jk。最佳抓取姿势 G\_j\_∗ = argmaxj C\_j 在从 P\_Depth 推断出的局部几何约束内选择最大化空间一致性的姿势。请注意，物体放置的姿态和独立动作 A 的最终姿态可以用类似的方法生成。

对于动作执行，机械臂的轨迹采用 Denavit-Hartenberg 运动学计算。插值路点经过优化，以在接近过程中保持末端执行器的方向约束，确保在每一步实时感知之后，整个任务中的运动平稳流畅。通过闭环方式，机器人可以有效地调整其操作目标以适应动态场景中的任务。

B. 灵巧操作姿态生成

平行夹持器的抓取感知方法在简化抓取合成方面取得了显著进展，这些进展可以作为灵巧操作的基础先验知识。具体而言，我们通过运动学重定向将平行夹持器的抓取方案迁移到灵巧手。这种方法可以利用现有的感知框架，同时有效地适应灵巧手更高的自由度，从而提高成功率，尤其是在抓取形状奇特的物体（例如杨桃）时。

抓取方案定义为 G = {t, R, w}，其中 t ∈ R3 是笛卡尔坐标系中的抓取中心，R ∈ SO(3) 是旋转矩阵，w 是成功抓取的夹持器宽度。一旦达到所需姿态，灵巧手的力传感模块会促使所有手指同时闭合，并持续闭合直至施加的力达到最大阈值 F\_max。

为了将 G 应用于灵巧抓取，我们需要确定从法兰框架 E 到相应灵巧手框架 H 的校准矩阵。首先，UR5 在基座框架 B 中的末端姿态可以通过以下公式计算：

其中，BCT 和 HET 分别为眼-手校准矩阵和手-末端校准矩阵。此外，手在相机框架 C 中的抓取姿态表示为：

由此，BET 可以通过公式 (5) 计算。

为了获得精确的手-末端校准矩阵 HET ，我们手动移动灵巧手的姿态并进行微调，以确保拇指和中指的当前姿态与 3D 点云中显示的双指夹持器的姿态相对应。然后记录当前的 HBT，最后，可以通过公式 (4) 计算校准矩阵。所有上述框架均在图 3 中进行了演示，以便于理解灵巧抓取姿态的生成过程。

图 3. RoboDexVLM 的机器人操控系统设置。机器人操控器 (UR5) 旨在抓取和操控桌面上的物体，并使用灵巧手 (Inspire Hand) 作为末端执行器，并通过安装在灵巧手上的 RGB-D 摄像头 (RealSense D435i) 产生感知，与抽屉或其他类型的容器进行交互。图中分别显示了基座 {B}、手 {H}、末端执行器 {E} 和摄像头 {C} 的坐标。

C. 操控过程中的故障恢复策略

为了确保对执行错误的鲁棒性，RoboDexVLM 采用双层恢复机制。每次技能执行后，都会使用基于深度的变化检测和灵巧手所有手指的位置反馈来验证是否成功。如果发生失败，例如抓取失败，系统会构建一个反思提示，其中包含检测到的错误 E\_error、当前场景状态 P\_RGB 以及先前技能尝试历史记录 O\_history ⊂ Oτ 的详细信息。VLM 使用 {Rτ+1, Oτ+1, Iτ+1} = T (K (Hreflect)) 处理此提示，以建议调整后的技能序列。例如，如果物体在抓取尝试过程中滑落，系统可能会插入 HandRot 基元来重新调整物体方向，使其更牢固地抓握。为了防止无限循环，系统会从上次成功的技能继续执行，并将每个任务的恢复尝试次数限制为三次。第四部分中展示的实验结果表明，该策略显著提高了任务成功率，尤其是在长时域操控场景中。

IV.实验分析

A. 环境设置

实验在为机器人操作任务设计的真实环境中进行。如图 3 所示，该设置包括一个配备 Inspire 五指灵巧手的 UR5 机械臂、一个用于物体检测和场景分析的 Intel RealSense D435i RGB-D 摄像头，以及一个包含各种不同形状、大小和纹理物体的工作空间，以测试 RoboDexVLM 框架的多功能性。我们选择 GPT-4o [30] 作为灵巧机器人操作的基础模型。该系统可在不同的桌面布局下运行，以确保在不同场景下都能保持稳健的性能。所有计算和模型预测均在配备 RTX 3080Ti GPU 和 12 GB 显存的工作站上运行，以满足实时处理需求。

B. 零样本灵巧操作的有效性

我们通过演示不同开放词汇任务的中间过程，定性评估了 RoboDexVLM 的灵巧操作性能。

在将水果放入带盖盒子的过程中，要求机器人能够使用自然语言命令准确识别并抓取盒盖的把手。如图 4 第一列所示，指定的对象（盒盖把手）被正确遮盖。RoboDexVLM 利用上面生成的 B\_img，根据图 4 第四行中并联夹持器的抓取感知，生成了一个正常的抓取姿态。此外，使用第三部分 B 节中提出的灵巧姿态生成算法，灵巧手能够以类似人类的抓取姿势抓取盒盖。

图 4. 灵巧抓取姿态生成过程。用于分割的对象名称位于图的顶部。抓取姿势区域图像中的蓝色锚点表示抓取感知结果。

相反，如图 4 第二列所示，当执行开放词汇任务“将杨桃放入篮子”时，VLM 会针对图 4 第三行所示的精确物体掩码生成命令 E\_lang = carambola。在带掩码的 RGB 图像中，文本描述的物体的语义分割掩码也进行了相应的标记。如图 4 第四行所示，杨桃的不规则形状使其在使用平行夹持器抓取时容易受损，因为两个表面之间的垂直接触很容易损害水果的完整性。相比之下，转移后的灵巧抓取姿势为手部建立了一个包络面，使水果能够以半圆形的方式被握住。因此，这种抓取配置不仅减轻了平行夹持器垂直接触面的过大压力，还增强了抓取的稳定性。

此外，图 5 还展示了两个长时域操作任务的关键帧。在图 5(a) 中，机器人被指示将所有水果放入带盖的篮子中。VLM 生成一系列动作，包括先将盖子放在一边，再将所有水果放入打开的篮子中，这遵循了针对开放词汇任务定制的逻辑推理过程。图 5(b) 展示了另一项具有挑战性的操作任务“打开抽屉并取出里面的物品”。机器人首先检测并抓住抽屉的把手，然后拉开抽屉以识别存放在其中的物品。一旦检测到物品，机器人操纵器就会采取轻柔灵巧的手势将其取出并放置在橱柜表面上。

总而言之，RoboDexVLM 操作系统在处理纤细或形状奇特的物体方面表现出色，展现了其先进的灵活性和自适应操作机制。在长时域任务中展示的灵巧操作能力，为理解类人抓取以及任务执行的顺序提供了宝贵的见解。

图 5. 长视野灵巧操作演示。RoboDexVLM 框架的输入是一句描述待完成任务的句子。相关技能会自动调用，与开放词汇任务中的物体进行交互。相关视频可在项目页面查看。(a) 任务“将水果放入篮子”的关键帧。(b) 任务“打开抽屉并取出其中的物品”的关键帧。

C. 比较与简化研究

本研究旨在评估基于语言的开放世界物体分割模块相较于传统预定义和训练好的物体检测方法，在开放词汇任务的基础能力方面的优势。具体而言，本研究考察了基于 LangSAM [4] 在 RoboDexVLM 框架中实现的开放世界物体分割模块的贡献，并将其性能与经典物体检测方法 YOLOv11 [1] 进行了比较。评估涵盖了一系列不同的开放世界物体检测任务，这些任务具有不同的颜色、空间和大小属性。如表一所示，在相同的环境条件下，我们的语言驱动方法在所有评估任务场景中均展现出更高的成功率。尽管推理时间显著缩短，但结果揭示了 YOLO 检测流程在面对开放词汇设置中固有的细微操作要求时存在严重局限性。虽然两种方法都面临着类似的计算硬件限制，但我们的方法实现了近乎完美的成功率，而基准分数低于 55%。这种差异源于两个关键因素：语义基础保真度和属性推理。首先，基于语言的查询（例如“较小的杨桃”）能够实现像素级精确定位，而传统的边界框则无法实现。其次，关系描述符的显式建模消除了静态类别检测器所需的启发式位置过滤，从而减少了多对象交互过程中的级联错误。这些发现最终证实，从严格的分类法过渡到基于语言的分割范式，从根本上提升了机器人系统的操作范围，这是在非结构化环境中部署通用机械手的先决条件。

此外，我们通过性能指标以及与不可恢复方法的比较来评估恢复机制的有效性，并重点关注其对机器人任务可靠性的影响，如表二所示。结果表明，该恢复机制显著提升了任务性能。对于任务 1，单个水果语义排序（抓取失败），启用恢复机制的 RoboDexVLM 的成功率为 96.67%，而未启用恢复机制的成功率为 90.00%。成功率的略微提升，加上 31.5 ± 2.4 秒和 30.5 ± 2.3 秒的可比执行时间，表明即使是相对简单的任务也能受益于恢复机制的纠错能力，从而允许进行微小调整以提高整体可靠性。相比之下，任务 2，单个水果语义排序（物体位置改变），则展示了恢复机制更显著的影响。在执行时间相当的情况下，成功率从未启用恢复机制的 20.00% 提高到启用恢复机制的 96.67%，反映出在干扰条件下任务执行的显著改善。对于最后一类任务，即多对象排列，使用恢复机制的成功率比非恢复机制高 26.66%，这意味着对于长序列和复杂的任务，反思和从失败中重新规划是尤其必要的。所有任务的结果都清楚地表明，虽然恢复机制在更复杂的情况下可能会增加执行时间，但它最终提高了机器人动作的可靠性和成功率。效率和有效性之间的这种权衡强调了结合恢复策略的重要性，特别是对于需要更高精度和适应性的任务。此外，表 III 在不同开放词汇任务描述中进一步评估了记忆存储的有效性。对于简单的原子任务，没有记忆的平均成功率比使用记忆的版本低约 25.00%。随着任务复杂性和技能序列长度的增加，记忆模块的优势越来越明显。值得注意的是，在处理涉及打开盖子操作的任务时，没有记忆检索的版本的成功率仅为 20.00%，而 RoboDexVLM 的完整版本成功率可达 85.00%。

D. RoboDexVLM 的鲁棒性与风险分析

为了评估 RoboDexVLM 框架的鲁棒性，我们对表三所示的三类复杂度不同的开放词汇任务进行了定量实验。结果揭示了任务结构、推理效率和物理执行可靠性之间的关键关系。

表三前三行中的单水果语义排序任务表现出卓越的可靠性，30 次试验的成功率超过 90.00%，这得益于高效的推理和执行。这与我们的假设相符，即原子对象操作可以最大限度地减少累积不确定性。相比之下，表三第四行和第五行中与从抽屉中取出和放入物品相关的任务的成功率略有下降，这归因于复合挑战。与第一类任务相比，推理时间的延长（28.5 ± 7.3 秒）进一步反映了任务规划更全面、更长的推理过程。最后，虽然第三类任务——长视界多物体摆放——需要更长的推理时间，但由于抓握篮子的垂直把手比抓握安全裕度较小的水平把手更容易，其成功率与抽屉操作相当。最后，随着任务复杂度的提高，技能顺序 Oτ 的推理时间逐渐增加，如表三从上到下所示。这一观察结果与推理时间和执行时间的逐步增加一致。

五、结论

本文提出了 RoboDexVLM，这是一个新颖的灵巧操作框架，它将动态更新的变量存储机制与交互原语相结合，以应对开放词汇和长视界任务的挑战。通过将 VLM 与模块化技能库和测试时间自适应功能统一起来，我们的系统在从原子对象操作到复杂的多阶段操作等各种场景中展现出强大的适应性。关键创新包括：分层恢复机制，可缓解复杂或长期任务中的级联错误；基于语言的分割范式，可实现基于属性的精确物体灵巧抓取感知；以及闭环感知-动作流水线，专为现实世界灵巧操控部署而优化。RoboDexVLM 通过可重用原语将任务规划与底层控制分离，降低了非专业用户通过自然语言命令编写复杂灵巧机器人行为的门槛。未来的工作将侧重于将此范式扩展到动态多智能体协作场景，并使用因果推理模型增强故障预测。这项研究标志着我们朝着能够以最小的重构工作量可靠运行的通用操控系统迈出了关键一步。