# OmniManip: Towards General Robotic Manipulation via Object-Centric Interaction Primitives as Spatial Constraints

# OmniManip:通过以对象为中心的交互原语作为空间约束实现通用机器人操作

Mingjie Pan , Jiyao Zhang , Tianshu , Yinghao Zhao , Wenlong Gao , Hao Dong

潘明杰 ,张佳瑶 ,田舒 ,赵英豪 ,高文龙 ,董浩

CFCS, School of CS, Peking University PKU-AgiBot Lab AgiBot

北京大学计算机学院CFCS 北京大学AgiBot实验室 AgiBot

https://omnimanip.github.io

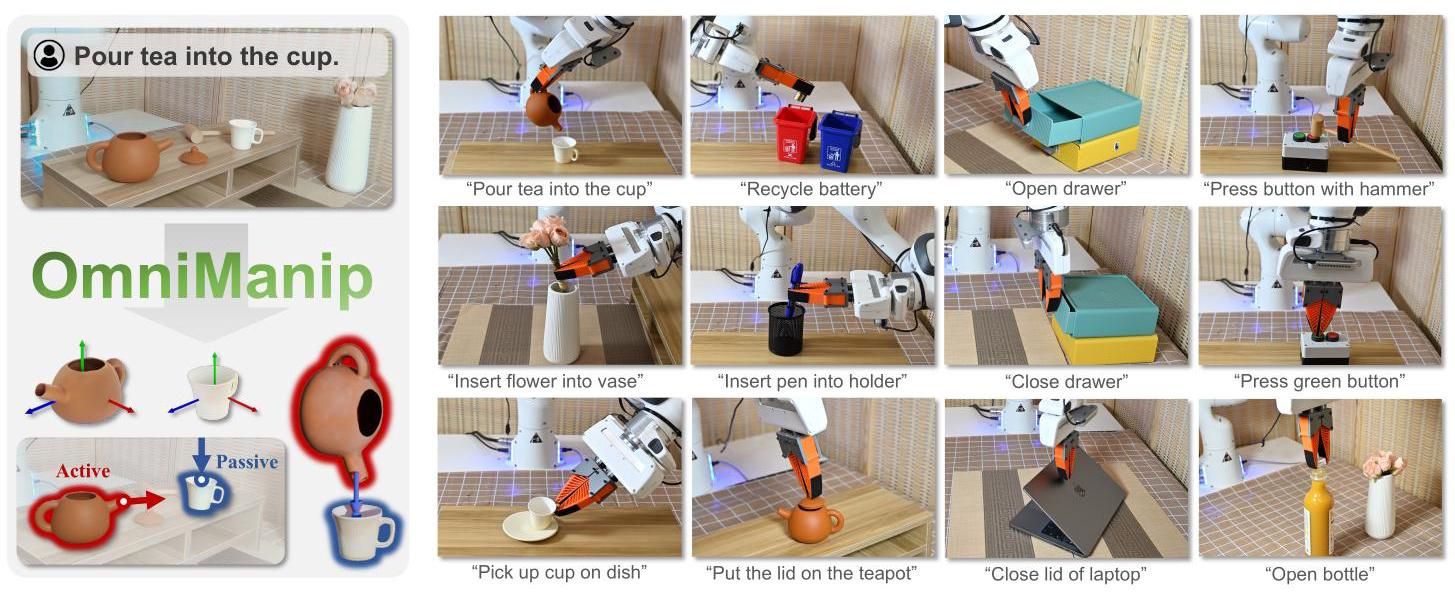


Figure 1. We proposed OmniManip, an open-vocabulary manipulation method that bridges the gap between the high-level reasoning of vision-language models (VLM) and the low-level precision, featuring closed-loop capabilities in both planning and execution.

图1. 我们提出了OmniManip，一种开放词汇的操作方法，弥合了视觉语言模型(VLM)的高层推理与低层精度之间的差距，具备规划和执行的双闭环能力。

# Abstract

# 摘要

The development of general robotic systems capable of manipulating in unstructured environments is a significant challenge. While Vision-Language Models(VLM) excel in high-level commonsense reasoning, they lack the fine-grained 3D spatial understanding required for precise manipulation tasks. Fine-tuning VLM on robotic datasets to create Vision-Language-Action Models(VLA) is a potential solution, but it is hindered by high data collection costs and generalization issues. To address these challenges, we propose a novel object-centric representation that bridges the gap between VLM’s high-level reasoning and the low-level precision required for manipulation. Our key insight is that an object’s canonical space, defined by its functional affor-dances, provides a structured and semantically meaningful way to describe interaction primitives, such as points and directions. These primitives act as a bridge, translating VLM’s commonsense reasoning into actionable 3D spatial constraints. In this context, we introduce a dual closed-loop, open-vocabulary robotic manipulation system: one loop for high-level planning through primitive resampling, interaction rendering and VLM checking, and another for low-level execution via pose tracking. This design ensures robust, real-time control without requiring VLM fine-tuning. Extensive experiments demonstrate strong zero-shot generalization across diverse robotic manipulation tasks, highlighting the potential of this approach for automating large-scale simulation data generation.

开发能够在非结构化环境中操作的通用机器人系统是一个重大挑战。虽然视觉语言模型(VLM)在高层常识推理方面表现出色，但它们缺乏精确操作任务所需的细粒度3D空间理解能力。在机器人数据集上微调VLM以创建视觉语言动作模型(VLA)是一种潜在的解决方案，但它受到高数据收集成本和泛化问题的阻碍。为了解决这些挑战，我们提出了一种新颖的以对象为中心的表示方法，弥合了VLM的高层推理与操作所需的低层精度之间的差距。我们的关键见解是，对象的功能性空间(由其功能可供性定义)提供了一种结构化和语义上有意义的方式来描述交互原语，如点和方向。这些原语充当桥梁，将VLM的常识推理转化为可操作的3D空间约束。在此背景下，我们引入了一种双闭环、开放词汇的机器人操作系统:一个循环通过原语重采样、交互渲染和VLM检查进行高层规划，另一个循环通过 姿态跟踪进行低层执行。这种设计确保了稳健的实时控制，而无需VLM微调。大量实验证明了在各种机器人操作任务中的强零样本泛化能力，突显了这种方法在自动化大规模仿真数据生成方面的潜力。

# 1. Introduction

# 1. 引言

Developing a general robotic manipulation system has long been a challenging task, primarily due to the complexity and variability of real-world [26, 47, 48]. Inspired by the rapid advancements in Large Language Models (LLM)[1,42] and Vision-Language Models (VLM) [25, 28, 34, 54], which leverage vast amounts of internet data to acquire rich commonsense knowledge, researchers have recently turned attention to exploring their application in robotics[14, 53]. Most existing works focus on utilizing this knowledge for high-level task planning, such as semantic reasoning . Despite these advances, current VLMs, primarily trained on extensive visual data, lack the spatial understanding ability necessary for precise, low-level manipulation tasks. This limitation poses challenges in manipulations within unstructured environments.

开发通用机器人操作系统长期以来一直是一项具有挑战性的任务，主要是由于现实世界的复杂性和多变性[26, 47, 48]。受到大语言模型(LLM)[1,42]和视觉语言模型(VLM)[25, 28, 34, 54]快速发展的启发，这些模型利用大量互联网数据获取丰富的常识知识，研究人员最近开始探索它们在机器人学中的应用[14, 53]。大多数现有工作集中在利用这些知识进行高层任务规划，如语义推理 。尽管取得了这些进展，当前的VLM主要是在广泛的 视觉数据上训练的，缺乏精确、低层操作任务所需的 空间理解能力。这一限制在非结构化环境中的操作中带来了挑战。

[[1]](#footnote-29)

One approach to overcoming this limitation is to fine-tune VLM on large-scale robotic datasets, transforming them into VLA [2,3,8,19]. However, this faces two major challenges: 1) acquiring diverse, high-quality robotic data is costly and time-consuming, and 2) fine-tuning VLM into VLA results in agent-specific representations, which are tailored to specific robots, limiting their generalizability. A promising alternative is to abstract robotic actions into interaction primitives (e.g., points or vectors) and leverage VLM reasoning to define the spatial constraints of these primitives, while traditional planning algorithms handle execution [13, 15, 27]. However, existing methods for defining and using primitives have several limitations: The process of generating primitive proposals is task-agnostic, which poses the risk of lacking suitable proposals. Additionally, relying on manually designed rules for post-processing proposals also introduces instability. This naturally leads to an important question: How can we develop more efficient and generalizable representations that bridge VLM high-level reasoning with precise, low-level robotic manipulation?

克服这一限制的一种方法是在大规模机器人数据集上微调VLM，将其转化为VLA [2,3,8,19]。然而，这面临两个主要挑战:1)获取多样化的高质量机器人数据成本高且耗时，2)将VLM微调为VLA会导致特定于代理的表示，这些表示针对特定机器人定制，限制了它们的泛化能力。一个有前途的替代方案是将机器人动作抽象为交互原语(如点或向量)，并利用VLM推理来定义这些原语的空间约束，而传统规划算法处理执行[13, 15, 27]。然而，现有定义和使用原语的方法有几个局限性:生成原语建议的过程与任务无关，这可能导致缺乏合适的建议。此外，依赖手动设计的规则对建议进行后处理也引入了不稳定性。这自然引出了一个重要问题:我们如何开发更高效和可泛化的表示方法，以弥合VLM的高层推理与精确、低层的机器人操作之间的差距？

To address this challenge, we propose a novel object-centric intermediate representation incorporating interaction points and directions within an object’s canonical space. This representation bridges the gap between VLM’s high-level commonsense reasoning and precise 3D spatial understanding. Our key insight is that an object’s canonical space is typically defined based on its functional af-fordances. As a result, we can describe an object’s functionality in a more structured and semantically meaningful way within its canonical space. Meanwhile, recent advancements in universal object pose estimation [7, 55, 56] make it feasible to canonicalize a wide range of objects.

为了解决这一挑战，我们提出了一种新颖的以对象为中心的中间表示，该表示结合了对象规范空间内的交互点和方向。这种表示弥合了视觉语言模型(VLM)的高级常识推理与精确的三维空间理解之间的差距。我们的关键见解是，对象的规范空间通常基于其功能可供性定义。因此，我们可以在其规范空间内以更结构化和语义上有意义的方式描述对象的功能。同时，通用对象姿态估计的最新进展[7, 55, 56]使得对广泛对象进行规范化成为可能。

Specifically, we employ a universal 6D object pose estimation model [56] to canonicalize objects and describe their rigid transformations during interactions. In parallel, a single-view 3D generation network generates detailed object meshes . Within the canonical space, interaction directions are initially sampled along the object’s principal axes, providing a coarse set of interaction possibilities. Meanwhile, the VLM predicts interaction points. Subsequently, the VLM identifies task-relevant primitives and estimates the spatial constraints between them. To address the hallucination issue in VLM reasoning, we introduce a self-correction mechanism through interaction rendering and primitive resampling that enables closed-loop reasoning. Once the final strategy is determined, actions are computed through constrained optimization, with pose tracking ensuring robust, real-time control in a closed-loop execution phase. Our method offers several key advantages: 1) Efficient and Effective Interaction Primitive Sampling: By leveraging the object’s canonical space, our approach enables efficient and effective sampling of interaction primitives, enhancing the system’s reasoning capabilities. 2) Dual Closed-Loop, Open-Vocabulary Robotic Manipulation System: Benefiting from the proposed object-centric intermediate representation, our method implements a dual closed-loop system. The rendering and resampling process drives a reasoning loop for decision-making, while pose tracking ensures a closed loop for action execution.

具体而言，我们采用了一种通用的6D对象姿态估计模型[56]来规范化对象并描述其在交互过程中的刚性变换。同时，单视图三维生成网络生成详细的对象网格 。在规范空间内，交互方向最初沿对象的主轴进行采样，提供一组粗略的交互可能性。与此同时，VLM预测交互点。随后，VLM识别与任务相关的基元并估计它们之间的空间约束。为了解决VLM推理中的幻觉问题，我们通过交互渲染和基元重采样引入了一种自我校正机制，实现了闭环推理。一旦确定了最终策略，通过约束优化计算动作，姿态跟踪确保在闭环执行阶段实现稳健的实时控制。我们的方法具有几个关键优势:1)高效且有效的交互基元采样:通过利用对象的规范空间，我们的方法能够高效且有效地采样交互基元，增强了系统的推理能力。2)双闭环、开放词汇的机器人操作系统:得益于提出的以对象为中心的中间表示，我们的方法实现了双闭环系统。渲染和重采样过程驱动决策推理循环，而姿态跟踪确保动作执行的闭环。

In summary, our contributions are threefold:

总之，我们的贡献有三点:

* We propose a novel object-centric interaction representation that bridges the gap between VLM’s high-level commonsense reasoning and low-level robotic manipulation.
* 我们提出了一种新颖的以对象为中心的交互表示，弥合了VLM的高级常识推理与低级机器人操作之间的差距。
* To the best of our knowledge, we are the first to present a planning and execution dual closed-loop open-vocabulary manipulation system without VLM fine-tuning.
* 据我们所知，我们是第一个提出无需VLM微调的规划和执行双闭环开放词汇操作系统的团队。
* Extensive experiments demonstrate our method’s strong zero-shot generalization across diverse manipulation tasks, and we also highlight its potential for automating robotic manipulation data generation.
* 大量实验表明，我们的方法在各种操作任务中表现出强大的零样本泛化能力，并且我们还强调了其在自动化机器人操作数据生成方面的潜力。

# 2. Related Work

# 2. 相关工作

Foundation Models For Robotics The emergence of foundation models has significantly influenced the field of robotics[11, 18, 51], particularly in the application of vision-language models[1, 4, 12, 23, 28, 50], which excel in environment understanding and high-level commonsense reasoning. These models demonstrate the potential for controlling robots to perform general tasks in novel and unstructured environments. Some studies have fine-tuned VLM on robotics datasets to create VLA models that output robotic trajectories, but these efforts are limited by the high cost of data collection and issues with generalization. Other approaches attempt to extract operation primitives using visual foundation models , which are then used as visual or language prompts for VLM to perform high-level commonsense reasoning, combined with motion planners for low-level control. However, these methods are constrained by the ambiguity of compressing primitives into the 2D images or 1D text required by VLM and the hallucination tendencies of VLM themselves, making it difficult to ensure that the high-level plans generated by VLM are accurate. In this work, we demonstrate Omni-Manip’s unique advantages in addressing these challenges, particularly in fine-grained 3D understanding and mitigating large model hallucinations.

机器人基础模型 基础模型的出现显著影响了机器人领域[11, 18, 51]，特别是在视觉语言模型的应用中[1, 4, 12, 23, 28, 50]，这些模型在环境理解和高级常识推理方面表现出色。这些模型展示了控制机器人在新颖和非结构化环境中执行通用任务的潜力。一些研究 在机器人数据集上对VLM进行了微调，以创建输出机器人轨迹的VLA模型，但这些努力受到数据收集成本高和泛化问题的限制。其他方法尝试使用视觉基础模型提取操作基元 ，然后将其作为视觉或语言提示供VLM执行高级常识推理，并结合运动规划器 进行低级控制。然而，这些方法受到将 基元压缩为VLM所需的二维图像或一维文本的模糊性以及VLM本身的幻觉倾向的限制，难以确保VLM生成的高级计划是准确的。在这项工作中，我们展示了Omni-Manip在解决这些挑战中的独特优势，特别是在细粒度三维理解和大模型幻觉缓解方面。

Representations for Manipulation Structural representations determine the capabilities and effectiveness of manipulation methods. Among various types of representations, keypoints are a popular choice due to their flexibility, generalization, and ability to model variability [32, 35, 36, 46]. However, these keypoints-based methods require manual task-specific annotations to generate actions. To enable zero-shot open-world manipulation, studies such as have transformed keypoints into visual prompts for VLM, facilitating the automatic generation of high-level planning results. Despite their advantages, keypoints can be unstable; they struggle under occlusion and pose challenges in the extraction and selection of specific keypoints. Another common representation is the pose, which efficiently defines long-range dependencies between objects for manipulation and offers a degree of robustness to occlusion [16, 17, 44, 45]. However, these methods necessitate prior modeling of geometric relationships and, due to the sparse nature of poses, cannot provide fine-grained geometry. This limitation can lead to failures in manipulation strategies across different objects due to intra-class variations. To address these issues, OmniManip combines the fine-grained geometry of keypoints with the stability of the 6D pose. It automatically extracts detailed functional points and directions within the canonical coordinate system of objects using VLM, enabling precise manipulation.

表示方法对操作能力的影响 结构表示方法决定了操作方法的能力和有效性。在各种表示方法中，关键点因其灵活性、泛化能力和对变化的建模能力而成为热门选择[32, 35, 36, 46]。然而，这些基于关键点的方法需要手动进行任务特定的标注以生成动作。为了实现零样本开放世界操作，诸如 的研究将关键点转化为视觉语言模型(VLM)的视觉提示，从而促进高级规划结果的自动生成。尽管关键点具有优势，但它们可能不稳定；在遮挡情况下表现不佳，并且在特定关键点的提取和选择上存在挑战。另一种常见的表示方法是 姿态，它有效地定义了对象之间的长程依赖关系，并提供了一定程度的遮挡鲁棒性[16, 17, 44, 45]。然而，这些方法需要对几何关系进行先验建模，并且由于姿态的稀疏性，无法提供细粒度的几何信息。这一限制可能导致由于类内变化而在不同对象上的操作策略失败。为了解决这些问题，OmniManip将关键点的细粒度几何与6D姿态的稳定性相结合。它利用VLM自动提取对象在规范坐标系中的详细功能点和方向，从而实现精确操作。

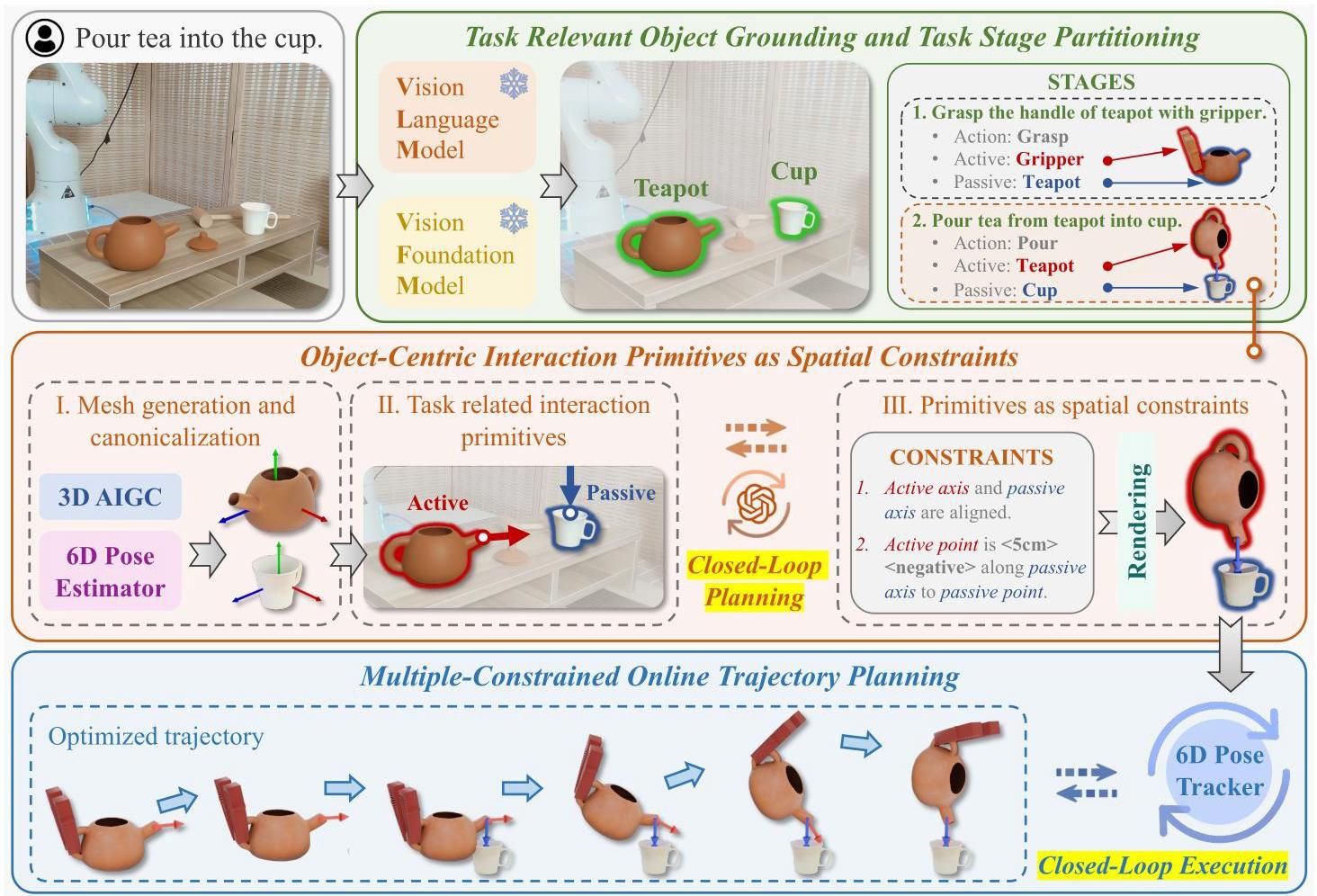


Figure 2. Overview framework. Given instruction and RGB-D observation marked by VFM, VLM firstly filters task-related objects and partitions the task into stages. For each stage, VLM extracts object-centric canonical interaction primitives as spatial constraints in a closed-loop manner. For execution, the trajectory is optimized by constraints and updated in a closed loop using a 6D Pose Tracker.

图2. 总体框架。给定由VFM标记的指令和RGB-D观测，VLM首先过滤与任务相关的对象，并将任务划分为多个阶段。对于每个阶段，VLM以闭环方式提取以对象为中心的规范交互原语作为空间约束。在执行过程中，轨迹通过约束进行优化，并使用6D姿态跟踪器在闭环中更新。

# 3. Method

# 3. 方法

Here we discuss: (1) How do we formulate robotic manipulation via interaction primitives as spatial constraints(Sec. 3.1)? (2) How to extract canonical interaction primitives in a generic and open vocabulary way (Sec. 3.2)? (3) Why can OmniManip achieve a dual closed-loop system (Sec. 3.3)?

在此我们讨论:(1) 如何通过交互原语作为空间约束来制定机器人操作(第3.1节)？(2) 如何以通用和开放词汇的方式提取规范交互原语(第3.2节)？(3) 为什么OmniManip能够实现双闭环系统(第3.3节)？

# 3.1. Manipulation with Interaction Primitives

# 3.1. 使用交互原语进行操作

In our formulation, complex robotic tasks are decomposed into stages, each defined by object interaction primitives with spatial constraints. This structured approach allows for the precise definition of task requirements and facilitates the execution of complex manipulation tasks. In this section, we detail how interaction primitives serve as the foundation for spatial constraints, enabling robust manipulation.

在我们的公式中，复杂的机器人任务被分解为多个阶段，每个阶段由具有空间约束的对象交互原语定义。这种结构化方法允许精确定义任务需求，并促进复杂操作任务的执行。在本节中，我们详细介绍了交互原语如何作为空间约束的基础，从而实现鲁棒的操作。

Task Decomposition. As shown in Figure 2, given a manipulation task (e.g., pouring tea into a cup), we first utilize GroundingDINO[30] and SAM[20], two Visual Foundation Models (VFMs), to mark all foreground objects in the scene like [49] as visual prompt. Subsequently, a VLM [1] is employed to filter task-relevant ob-

任务分解。如图2所示，给定一个操作任务 (例如，将茶倒入杯中)，我们首先利用GroundingDINO[30]和SAM[20]这两个视觉基础模型(VFMs)来标记场景中的所有前景对象，如[49]所示，作为视觉提示。随后，使用VLM[1]来过滤与任务相关的对象，并将任务分解为多个阶段 [latex1]，其中每个阶段[latex2]可以形式化为[latex3]，其中[latex4]表示要执行的动作(例如，抓取、倾倒)，[latex5]和[latex6]分别指发起交互的对象和被操作的对象。例如，在图2中，茶壶在抓取茶壶的阶段中是被动对象，而在将茶倒入杯子的阶段中，茶壶是主动对象，杯子是被动对象。

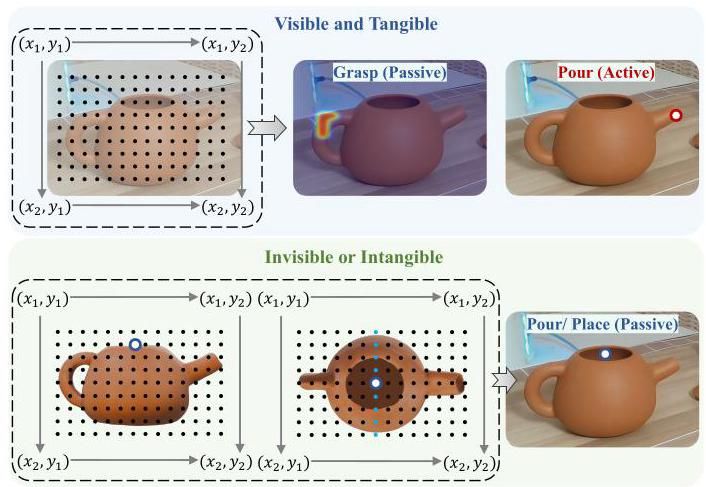


Figure 3. Interaction points generation.

图3. 交互点生成。

jects and decompose the task into multiple stages , where each stage can be formalized as , where represents the action to be performed (e.g., grasp, pour), and and refer to the object initiating the interaction and the object being acted upon, respectively. For example, in Figure 2, the teapot is the passive object in the stage of grasping the teapot while the teapot is the active object and the cup is passive in the stage of pouring tea into the cup.

对象并将任务分解为多个阶段 ，其中每个阶段 可以形式化为 ，其中 表示要执行的动作(例如，抓取、倾倒)， 和 分别指发起交互的对象和被操作的对象。例如，在图2中，茶壶在抓取茶壶的阶段中是被动对象，而在将茶倒入杯子的阶段中，茶壶是主动对象，杯子是被动对象。

Object-Centric Canonical Interaction Primitives. We propose a novel object-centric representation with canonical interaction primitives to describe how objects interact during manipulation tasks. Specifically, an object’s interaction primitives are characterized by its interaction point and direction in canonical space. The interaction point denotes a key location on the object where interaction occurs, while the interaction direction represents the primary axis relevant to the task. Together, these form the interaction primitive , encapsulating the essential intrinsic geometric and functional properties required to meet task constraints. These canonical interaction primitives are defined relative to their canonical space, remaining consistent across different scenarios, enabling more generalized and reusable manipulation strategies.

以对象为中心的规范交互原语。我们提出了一种新颖的以对象为中心的表示方法，使用规范交互原语来描述在操作任务中对象如何交互。具体来说，对象的交互原语由其规范空间中的交互点和交互方向来表征。交互点 表示对象上发生交互的关键位置，而交互方向 表示与任务相关的主要轴。这些共同构成了交互原语 ，封装了满足任务约束所需的基本内在几何和功能属性。这些规范交互原语是相对于其规范空间定义的，在不同场景中保持一致，从而实现更通用和可重用的操作策略。

Interaction Primitives with Spatial Constraints. At each stage , a set of spatial constraints governs the spatial relationships between the active and passive objects. These constraints are divided into two categories: distance constraints , which regulate the distance between interaction points, and angular constraints , which ensure proper alignment of interaction directions. Together, these constraints define the geometric rules necessary for precise spatial alignment and task execution. The overall spatial constraint for each stage is given by:

具有空间约束的交互原语。在每个阶段 ，一组空间约束 控制主动对象和被动对象之间的空间关系。这些约束分为两类:距离约束 ，用于调节交互点之间的距离；角度约束 ，用于确保交互方向的正确对齐。这些约束共同定义了精确空间对齐和任务执行所需的几何规则。每个阶段 的整体空间约束由以下公式给出:

Once the constraints have been defined, the task execution can be formulated as an optimization problem.

一旦定义了约束 ，任务执行就可以被表述为一个优化问题。

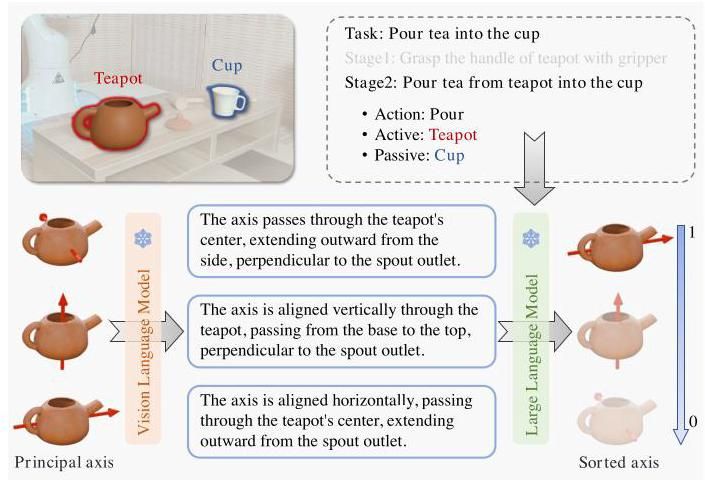


Figure 4. Interaction directions extraction.

图4. 交互方向提取。

# 3.2. Primitives and Constraints Extraction

# 3.2. 原语与约束提取

In this section, we detail the process of extracting interaction primitives and their spatial constraints for each stage. As illustrated in Figure 2, we first obtain 3D object meshes for both the task-relevant active and passive objects via single-view 3D generation [29, 40, 57], followed by pose estimation with Omni6DPose[56] for object canonicalization. Next, we extract task-relevant interaction primitives and their corresponding constraints.

在本节中，我们详细介绍了为每个阶段提取交互原语及其空间约束 的过程。如图2所示，我们首先通过单视图3D生成[29, 40, 57]获取任务相关的主动和被动对象的3D网格，然后使用Omni6DPose[56]进行姿态估计以实现对象规范化。接下来，我们提取任务相关的交互原语及其相应的约束。

Grounding Interaction Point. As shown in Figure 3, interaction points are categorized as Visible and Tangible (e.g., a teapot handle) or Invisible or Intangible (e.g., the center of its opening). To enhance VLM for interaction points grounding, SCAFFOLD [22] visual prompting mechanism is employed, which overlays a Cartesian grid onto the input image. Visible points are directly localized in the image plane, while invisible points are inferred through multi-view reasoning based on proposed canonical object representations, as illustrated in Figure 3. Reasoning begins from the primary viewpoint, with ambiguities resolved by switching to an orthogonal view. This approach enables more flexible and reliable interaction point grounding. For tasks like grasping, heatmaps are generated from multiple interaction points, improving the robustness of the grasping model.

交互点定位。如图3所示，交互点分为可见且可触摸的(例如茶壶把手)或不可见或不可触摸的(例如其开口中心)。为了增强视觉语言模型(VLM)在交互点定位中的应用，采用了SCAFFOLD[22]视觉提示机制，该机制将笛卡尔网格叠加到输入图像上。可见点直接在图像平面中定位，而不可见点则通过基于提出的规范对象表示的多视图推理来推断，如图3所示。推理从主视角开始，通过切换到正交视图来解决歧义。这种方法使得交互点定位更加灵活和可靠。对于抓取等任务，从多个交互点生成热图，提高了抓取模型的鲁棒性。

Sampling Interaction Direction. In the canonical space, the principal axes of an object are often functionally relevant. As illustrated in Figure 4, we treat the principal axes as candidate interaction directions. However, assessing the relevance of these directions to the task is challenging due to the limited spatial understanding of the current VLM. To address this, we propose a VLM caption and LLM scoring mechanism: first, we use the VLM to generate semantic descriptions for each candidate axis, and then employ a LLM to infer and score the relevance of these descriptions to the task. This process results in an ordered set of candidate directions that are most aligned with the task requirements.

交互方向采样。在规范空间中，对象的主轴通常与功能相关。如图4所示，我们将主轴视为候选交互方向。然而，由于当前VLM的空间理解有限，评估这些方向与任务的相关性具有挑战性。为了解决这个问题，我们提出了一个VLM描述和LLM评分机制:首先，我们使用VLM为每个候选轴生成语义描述，然后使用LLM推断并评分这些描述与任务的相关性。这个过程产生了一组与任务要求最匹配的有序候选方向。

Ultimately, the interaction primitives with constraints are generated with VLM, yielding an ordered list of constrained interaction primitives for each stage , denoted as .

最终，使用VLM生成带有约束的交互原语，为每个阶段 生成一个有序的约束交互原语列表，表示为 。

# 3.3. Dual Closed-Loop System

# 3.3. 双闭环系统

As outlined in Section 3.2, we obtain the interaction primitives of the active and passive objects, denoted as and , respectively, along with the spatial constraints that define their spatial relationships. However, this is an open-loop inference, which inherently limits the robustness and adaptability of the system. These limitations arise primarily from two sources: 1) the hallucination effect in large models, and 2) the dynamic nature of real-world environments. To overcome these challenges, we propose a dual closed-loop system, as illustrated in Figure 2.

如第3.2节所述，我们获得了主动和被动对象的交互原语，分别表示为 和 ，以及定义它们空间关系的空间约束 。然而，这是一个开环推理，这固有地限制了系统的鲁棒性和适应性。这些限制主要来自两个来源:1)大模型中的幻觉效应，2)现实世界环境的动态性。为了克服这些挑战，我们提出了一个双闭环系统，如图2所示。

Algorithm 1 Self-Correction Algorithm via RRC

算法1 通过RRC的自校正算法

Input: Task , Stage , Initial List of Primitives with Con-

straints

Output: Successful Constraints or Task Failure

1: Steps , refine False

while maxSteps do

Render:

Check: state

if state ’Refine’ and refine False then

Resample: Update Resample

, refine True

else if state ’Success’ then

return

end if

end while

return Task Failed

Closed-loop Planning. To improve the accuracy of interaction primitives and mitigate hallucination issues in the VLM, we introduce a self-correction mechanism based on Resampling, Rendering, and Checking (RRC). This mechanism uses real-time feedback from a visual language model (VLM) to detect and correct interaction errors, ensuring precise task execution. The RRC process consists of two stages: the initial phase and the refinement phase. The overall RRC mechanism is outlined in Algorithm 1. In the initial phase, the system evaluates the interaction constraints defined in Section 3.2, which specify the spatial relationships between active and passive objects. For each constraint , the system renders an interaction image based on the current configuration and submits it to the VLM for validation. The VLM returns one of three outcomes: success, failure, or refinement. If success, the constraint is accepted, and the task proceeds. If failure, the next constraint is evaluated. If refinement, the system enters the refinement phase for further optimization. In the refinement phase, the system performs fine-grained resampling around the predicted interaction direction to correct misalignments between the functional and geometric axes of objects. The system uniformly samples six refined directions around and evaluates them.

闭环规划。为了提高交互原语的准确性并减轻VLM中的幻觉问题，我们引入了一种基于重采样、渲染和检查(RRC)的自校正机制。该机制使用视觉语言模型(VLM)的实时反馈来检测和校正交互错误，确保任务的精确执行。RRC过程包括两个阶段:初始阶段和优化阶段。整个RRC机制在算法1中概述。在初始阶段，系统评估第3.2节中定义的交互约束 ，这些约束指定了主动和被动对象之间的空间关系。对于每个约束 ，系统根据当前配置渲染一个交互图像 并将其提交给VLM进行验证。VLM返回三种结果之一:成功、失败或优化。如果成功，则接受该约束并继续任务。如果失败，则评估下一个约束。如果优化，则系统进入优化阶段进行进一步优化。在优化阶段，系统在预测的交互方向 周围进行细粒度重采样，以校正对象功能轴和几何轴之间的偏差。系统在 周围均匀采样六个优化方向 并对其进行评估。

Closed-loop Execution. Once the interaction primitives and the corresponding spatial constraints are defined for each stage, the task execution can be formulated as an optimization problem. The objective is to minimize the loss function to determine the target pose of the end-effector. The optimization problem can be expressed as:

闭环执行。一旦为每个阶段定义了交互原语和相应的空间约束 ，任务执行就可以表述为一个优化问题。目标是最小化损失函数以确定末端执行器的目标姿态 。优化问题可以表示为:

where the constraint loss ensures that the action adheres to the task’s spatial constraints , and is defined as

其中约束损失 确保动作遵循任务的空间约束 ，并定义为

Here, measures the deviation between the current spatial relationship of the active object and the passive object from the desired constraint , while maps the end-effector pose to the active object’s pose. The collision loss prevents the end-effector from colliding with obstacles in the environment and is defined as

这里， 测量主动对象 和被动对象 的当前空间关系与期望约束 之间的偏差，而 将末端执行器姿态映射到主动对象的姿态。碰撞损失 防止末端执行器与环境中的障碍物碰撞，并定义为

where represents the distance between the end-effector and the obstacle , and is the minimum allowable safety distance. The path loss ensures smooth motion and is defined as

其中 表示末端执行器与障碍物 之间的距离， 是最小允许安全距离。路径损失 确保平滑运动，并定义为

where and represent the translational and rotational displacements of the end-effector, respectively, and and are weighting factors that balance the influence of translation and rotation. By minimizing these loss functions, the system dynamically adjusts the end-effector pose , ensuring successful task execution while avoiding collisions and maintaining smooth motion.

其中 和 分别表示末端执行器的平移和旋转位移， 和 是平衡平移和旋转影响的权重因子。通过最小化这些损失函数，系统动态调整末端执行器姿态 ，确保任务成功执行，同时避免碰撞并保持平滑运动。

While Equation 3 outlines how interaction primitives and their corresponding spatial constraints can be leveraged to optimize the executable end-effector pose, real-world task execution often involves significant dynamic factors. For instance, deviations in the grasp pose may result in unintended object movement during a grasping task. Moreover, in certain dynamic environments, the target object may be displaced. These challenges highlight the critical importance of closed-loop execution in handling such uncertainties. To address these challenges, our system leverages the proposed object-centric interaction primitives and directly employs an off-the-shelf 6D object pose tracking algorithm to continuously update the poses of both the active object and the passive object in real-time, as required in Equation 4. This real-time feedback allows for dynamic adjustments to the target pose of the end-effector, enabling robust and accurate closed-loop execution.

虽然公式3概述了如何利用交互原语及其相应的空间约束来优化可执行末端执行器的姿态，但现实世界中的任务执行通常涉及显著的动态因素。例如，抓取姿态的偏差可能导致在抓取任务中出现意外的物体移动。此外，在某些动态环境中，目标物体可能会发生位移。这些挑战凸显了闭环执行在处理此类不确定性中的关键重要性。为了解决这些挑战，我们的系统利用所提出的以物体为中心的交互原语，并直接采用现成的6D物体姿态跟踪算法，根据公式4的要求，实时更新主动物体 和被动物体 的姿态。这种实时反馈允许对末端执行器的目标姿态进行动态调整，从而实现稳健且准确的闭环执行。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tasks | VoxPoser |  | ReKep Auto | OmniManip(Ours) | |
| Closed-loop | Open-loop |
| Pour tea | 0/10 | 1/10 | 3/10 | 7/10 | 6/10 |
| Insert flower into vase | 0/10 | 4/10 | 2/10 | 6/10 | 4/10 |
| Insert the pen in holder | 0/10 | 4/10 | 3/10 | 7/10 | 5/10 |
| Recycle the battery | 6/10 | 5/10 | 7/10 |  | 6/10 |
| Pick up the cup on the dish | 3/10 | 2/10 | 9/10 | 8/10 | 7/10 |
| Fit the lid onto the teapot | 0/10 | 2/10 | 3/10 | 5/10 | 3/10 |
| Total | 15.0% | 30.0% | 45.0% | 68.3% | 51.7% |
| Open the drawer | 1/10 | 4/10 | - | 6/10 | 4/10 |
| Close the drawer | 3/10 | 3/10 | - | 8/10 | 6/10 |
| Hammer the button | 0/10 | 3/10 | - | 4/10 | 2/10 |
| Press the red button | 0/10 | 3/10 | - | 7/10 | 6/10 |
| Close the lid of the laptop | 4/10 | 3/10 | - | 6/10 | 4/10 |
| Open the jar | 2/10 | 0/10 | - | 6/10 | 5/10 |
| Total | 16.7% | 26.7% | - | 61.7% | 45.0% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 任务 | VoxPoser |  | ReKep Auto | OmniManip(我们的) | |
| 闭环 | 开环 |
| 倒茶 | 0/10 | 1/10 | 3/10 | 7/10 | 6/10 |
| 将花插入花瓶 | 0/10 | 4/10 | 2/10 | 6/10 | 4/10 |
| 将笔插入笔筒 | 0/10 | 4/10 | 3/10 | 7/10 | 5/10 |
| 回收电池 | 6/10 | 5/10 | 7/10 |  | 6/10 |
| 拿起盘子上的杯子 | 3/10 | 2/10 | 9/10 | 8/10 | 7/10 |
| 将盖子盖在茶壶上 | 0/10 | 2/10 | 3/10 | 5/10 | 3/10 |
| 总计 | 15.0% | 30.0% | 45.0% | 68.3% | 51.7% |
| 打开抽屉 | 1/10 | 4/10 | - | 6/10 | 4/10 |
| 关闭抽屉 | 3/10 | 3/10 | - | 8/10 | 6/10 |
| 敲击按钮 | 0/10 | 3/10 | - | 4/10 | 2/10 |
| 按下红色按钮 | 0/10 | 3/10 | - | 7/10 | 6/10 |
| 合上笔记本电脑的盖子 | 4/10 | 3/10 | - | 6/10 | 4/10 |
| 打开罐子 | 2/10 | 0/10 | - | 6/10 | 5/10 |
| 总计 | 16.7% | 26.7% | - | 61.7% | 45.0% |

Table 1. Quantitative results across 12 real-world manipulation tasks. The first six tasks focus on rigid object manipulation, while the latter involves articulated object manipulation. "-" indicates that the method can not handle this task due to its underlying principles.

表1. 12个现实世界操作任务的定量结果。前六个任务专注于刚性物体操作，而后六个任务涉及铰接物体操作。"-"表示该方法由于其基本原理无法处理此任务。

# 4. Experiment

# 4. 实验

In this section, we aim to answer the following questions: (1) To what extent does OmniManip perform effectively in open-vocabulary manipulation tasks across diverse real-world scenarios (Section 4.2)? (2) What role do the system’s critical features play in enhancing its overall performance (Section 4.3)? (3) How promising is OmniManip for automating the collection of robot manipulation trajectories to enable scalable imitation learning (Section 4.4)?

在本节中，我们旨在回答以下问题:(1) OmniManip在多样化的现实场景中的开放词汇操作任务中表现如何(第4.2节)？(2) 系统的关键特性在提升其整体性能中扮演了什么角色(第4.3节)？(3) OmniManip在自动化收集机器人操作轨迹以实现可扩展的模仿学习方面有多大的潜力(第4.4节)？

# 4.1. Experimental Setup

# 4.1. 实验设置

Hardware Configuration. Our experimental platform is built around a Franka Emika Panda robotic arm, with its parallel gripper’s fingers replaced by UMI fingers[6]. For perception, we employ two Intel RealSense D415 depth cameras. One camera is mounted at the gripper to provide a first-person view of the manipulation area, while the second camera is positioned opposite the robot to offer a third-person view of the workspace.

硬件配置。我们的实验平台基于Franka Emika Panda机械臂构建，其平行夹爪的手指被UMI手指[6]替换。为了感知，我们使用了两台Intel RealSense D415深度相机。一台相机安装在夹爪上，提供操作区域的第一人称视角，而第二台相机则放置在机器人对面，提供工作空间的第三人称视角。

Tasks and Metrics. As shown in Figure 1, We designed 12 tasks to evaluate models’ manipulation capabilities in real-world scenarios. Six of these involve rigid object manipulation (e.g., pour tea), while the others focus on articulated manipulation (e.g., open the drawer). These tasks cover a diverse set of objects and are intended to assess the models’ ability to generalize and adapt in complex environments. For each task, 10 trials were performed for each approach, and the success rate was recorded. After each trial, the object layout was reconfigured to ensure robust evaluation.

任务和指标。如图1所示，我们设计了12个任务来评估模型在现实场景中的操作能力。其中六个任务涉及刚性物体操作(例如倒茶)，而其他任务则专注于铰接操作(例如打开抽屉)。这些任务涵盖了多种物体，旨在评估模型在复杂环境中的泛化和适应能力。对于每个任务，每种方法进行了10次试验，并记录了成功率。每次试验后，物体布局都会重新配置，以确保评估的稳健性。

Baselines. We compare our approach with three baselines: 1) VoxPoser[14], which uses LLM and VLM to generate 3D value maps for synthesizing robot trajectories, excelling in zero-shot learning and closed-loop control; 2) , which introduces spatial constraints of object parts and combines with VLM to enable open-vocabulary manipulation; and 3) ReKep[15], which employs relational keypoint constraints and hierarchical optimization for real-time action generation from natural language instructions.

基线。我们将我们的方法与三个基线进行比较:1) VoxPoser[14]，它使用LLM和VLM生成3D值图以合成机器人轨迹，擅长零样本学习和闭环控制；2) ，它引入了物体部件的空间约束，并与VLM结合以实现开放词汇操作；3) ReKep[15]，它采用关系关键点约束和分层优化，从自然语言指令中实时生成动作。

Implement Details We use GPT-4O from OpenAI API as the vision-language model, leveraging a small set of interaction examples as prompts to guide the model’s reasoning for manipulation tasks. The specific prompts used are detailed in the appendix. We employ off-the-shelf models [10, 43] for 6-DOF universal grasping and utilize GenPose++[56] for universal pose estimation.

实现细节。我们使用OpenAI API中的GPT-4O作为视觉语言模型，利用一小部分交互示例作为提示，指导模型进行操作任务的推理。具体使用的提示在附录中详细说明。我们使用现成的模型[10, 43]进行6自由度通用抓取，并利用GenPose++[56]进行通用 姿态估计。

# 4.2. Open-Vocabulary Manipulation

# 4.2. 开放词汇操作

We conducted a comprehensive evaluation of OmniMa-nip on 12 open-vocabulary manipulation tasks, ranging from straightforward actions such as pick-and-place to more complex tasks involving object-object interactions with directional constraints and articulated object manipulation. As shown in Table 1, our method exhibits robust zero-shot generalization and superior performance across the board without task-specific training. This generalization capability can be attributed to the commonsense knowledge embedded in VLM, while the proposed efficient object-centric interaction primitives facilitate precise perception and execution. Additionally, we provide qualitative results in the appendix. OmniManip exhibits a substantial performance advantage over baseline methods, primarily due to two key factors: 1) the efficiency and stability of the proposed object-centric canonical interaction primitives, as further validated through extensive experiments in Section 4.3, and 2) the advanced dual closed-loop system for planning and execution. By incorporating a novel self-correction mechanism based on RRC, the system effectively mitigates hallucination issues of large models. As shown in Table 1, this closed-loop planning yields over a 15% improvement in performance for both rigid and articulated object manipulation tasks. A detailed qualitative analysis of the closed-loop reasoning and execution is provided in Section 4.3.

我们对OmniManip在12个开放词汇操作任务上进行了全面评估，从简单的拾取和放置动作到涉及方向约束的物体间交互和铰接物体操作的更复杂任务。如表1所示，我们的方法在没有任务特定训练的情况下表现出强大的零样本泛化能力和整体优越性能。这种泛化能力可以归因于VLM中嵌入的常识知识，而提出的高效以物体为中心的交互原语促进了精确的 感知和执行。此外，我们在附录中提供了定性结果。OmniManip在基线方法上表现出显著的性能优势，主要归因于两个关键因素:1) 提出的以物体为中心的规范交互原语的效率和稳定性，如第4.3节中的大量实验进一步验证；2) 先进的规划和执行双闭环系统。通过引入基于RRC的新型自校正机制，系统有效缓解了大模型的幻觉问题。如表1所示，这种闭环规划在刚性和铰接物体操作任务中的性能提升了超过15%。第4.3节提供了闭环推理和执行的详细定性分析。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| ReKep | 0/10 | 1/10 | 3/10 | 5/10 | 7/10 |
| OmniManip | 7/10 | 8/10 |  | 7/10 | 7/10 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| ReKep | 0/10 | 1/10 | 3/10 | 5/10 | 7/10 |
| OmniManip | 7/10 | 8/10 |  | 7/10 | 7/10 |

Table 2. Quantitative analysis of the impact of viewpoints on the performance, using ’Recycle the battery’ as a case study.

表2. 以“回收电池”为案例研究，定量分析视角对性能的影响。

# 4.3. Core Attributes of OmniManip

# 4.3. OmniManip的核心属性

Reliability of OmniManip. To effectively bridge VLM with low-level manipulation, reliable interaction primitives are crucial. We evaluate this across two key dimensions: stability and viewpoint consistency. Stability indicates the reliable extraction of task-relevant interaction primitives. As shown in Figure 5, ReKep extracts keypoint proposals through semantic clustering but lacks sensitivity to spatial geometry and task, making it challenging to generate sufficient task-relevant keypoints. CoPa extracts parts via explicit pixel segmentation, exhibiting high sensitivity to image texture and part shape. In contrast, OmniManip, an object-centric interaction primitive, samples interaction points in a canonical space aligned with the object’s functionality, ensuring both robustness and task-specific precision. Consistency of primitive extraction across varying viewpoints is critical to ensuring the stability of manipulation. Both ReKep and CoPa exhibit difficulties in this regard due to their reliance on sampling points directly from the object’s surface. Taking ReKep as an example, Figure 6 illustrates the planning results of ReKep and OmniManip for the ’Recycle battery’ task across different viewpoints. As shown, ReKep successfully identifies interaction points from a top-down view but fails under a frontal view, where the ideal target point is floating in the air. In contrast, OmniManip utilizes an object-centric primitive representation in a canonical space, ensuring viewpoint invariance. Table 2 presents the quantitative comparison, demonstrating that OmniManip’s performance is nearly invariant across varying viewpoints, whereas ReKep’s performance

OmniManip的可靠性。为了有效连接视觉语言模型(VLM)与低级操作，可靠的交互原语至关重要。我们从两个关键维度对此进行评估:稳定性和视角一致性。稳定性表示任务相关交互原语的可靠提取。如图5所示，ReKep通过语义聚类提取关键点建议，但对空间几何和任务缺乏敏感性，难以生成足够的任务相关关键点。CoPa通过显式像素分割提取部件，对图像纹理和部件形状表现出高敏感性。相比之下，OmniManip作为一种以对象为中心的交互原语，在与对象功能对齐的规范空间中采样交互点，确保了鲁棒性和任务特定的精确性。在不同视角下提取原语的一致性对于确保操作的稳定性至关重要。ReKep和CoPa在这方面都表现出困难，因为它们依赖于直接从对象表面采样点。以ReKep为例，图6展示了ReKep和OmniManip在不同视角下“回收电池”任务的规划结果。如图所示，ReKep在 俯视图中成功识别了交互点，但在 正面视图中失败，理想的目标点漂浮在空中。相比之下，OmniManip在规范空间中使用以对象为中心的原语表示，确保了视角不变性。表2展示了定量比较，表明OmniManip的性能在不同视角下几乎不变，而ReKep的性能

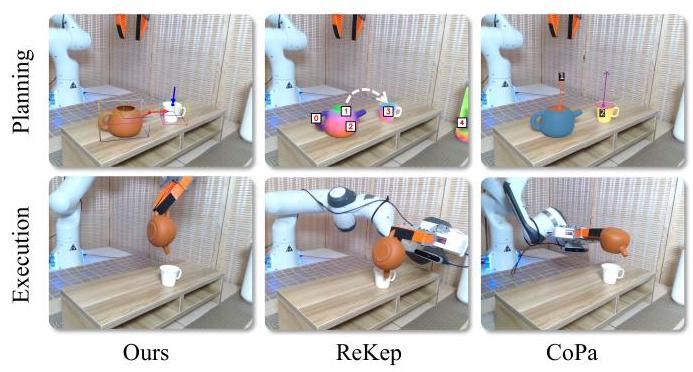


Figure 5. Stability analysis of interaction primitives. Visualization of planning and corresponding execution results across different methods, demonstrated using the ’Pour tea’ as a case study.

图5. 交互原语的稳定性分析。以“倒茶”为案例研究，展示不同方法的规划和相应执行结果的可视化。

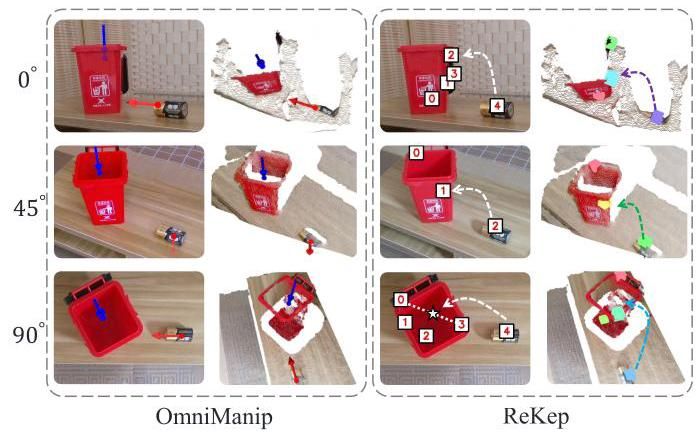


Figure 6. Qualitative analysis of the impact of viewpoints on the performance, using ’Recycle the battery’ as a case study.

图6. 以“回收电池”为案例研究，定性分析视角对性能的影响。

is significantly affected by changes in viewpoint.

显著受视角变化的影响。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sampling | Recycle Battery | | Pour Tea | |
| Suc. Rate | Iter. | Suc. Rate | Iter. |
| Uniform | 50% | 1.8 | 30% | 3.4 |
| OmniManip | 80% | 1.7 | 70% | 1.8 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 采样 | 回收电池 | | 倒茶 | |
| 成功率 | 迭代 | 成功率 | 迭代 |
| 均匀 | 50% | 1.8 | 30% | 3.4 |
| 全能操作 | 80% | 1.7 | 70% | 1.8 |

Table 3. Quantitative analysis of the primitive sampling efficiency.

表3. 原始采样效率的定量分析。

Efficiency of OmniManip. Interaction direction proposals in OmniManip are driven by a targeted sampling strategy. Compared with uniform sampling in , OmniManip samples along the principal axes of the object’s canonical space. Since the canonical space is aligned with the object’s functionality, this ensures both efficient and effective sampling. To evaluate this efficiency, we compared OmniMa-nip’s sampling strategy with uniform sampling in using two key metrics: the number of iterations and the corresponding task success rate. As shown in Table 3 OmniMa-nip not only requires fewer iterations but also achieves superior task performance, demonstrating that aligning the sampling process with the object’s functionality reduces sampling overhead while improving overall performance.

OmniManip的效率。OmniManip中的交互方向提议由目标采样策略驱动。与 中的均匀采样相比，OmniManip沿物体规范空间的主轴进行采样。由于规范空间与物体的功能对齐，这确保了采样既高效又有效。为了评估这种效率，我们使用两个关键指标将OmniManip的采样策略与 中的均匀采样进行了比较:迭代次数和相应的任务成功率。如表3所示，OmniManip不仅需要更少的迭代次数，而且实现了更优的任务性能，这表明将采样过程与物体的功能对齐可以减少采样开销，同时提高整体性能。

Closed-Loop Planning. In current methods, the planning component of VLM operates in an open-loop manner, meaning it cannot verify the correctness of the plan before execution. While ReKep achieves closed-loop control through point tracking, this only functions at the execution stage and does not provide feedback on the planning results generated by the VLM. In contrast, OmniManip introduces a unique self-correction mechanism via RRC, achieving closed-loop planning, which significantly reduces planning failures caused by VLM hallucinations, thereby offering more reliable planning. We report the results with closed-loop planning disabled in Table 1, where the task success rate decreases by over 15% in both rigid and articulated object manipulation tasks, demonstrating the effectiveness of the closed-loop planning approach. In Figure 7, we qualitatively illustrate the closed-loop planning results using the "Insert the pen in a holder" task as an example. It is evident that OmniManip can effectively pre-render the planning outcomes and achieve self-correction through the RRC process, thereby enabling closed-loop planning.

闭环规划。在当前的方法中，VLM的规划组件以开环方式运行，这意味着它在执行前无法验证计划的正确性。虽然ReKep通过点跟踪实现了闭环控制，但这仅在执行阶段起作用，并未对VLM生成的规划结果提供反馈。相比之下，OmniManip通过RRC引入了独特的自我校正机制，实现了闭环规划，这显著减少了由VLM幻觉引起的规划失败，从而提供了更可靠的规划。我们在表1中报告了禁用闭环规划的结果，其中刚性和铰接物体操作任务的任务成功率均下降了15%以上，证明了闭环规划方法的有效性。在图7中，我们以“将笔插入笔筒”任务为例，定性地展示了闭环规划结果。显然，OmniManip可以有效地预渲染规划结果，并通过RRC过程实现自我校正，从而实现闭环规划。

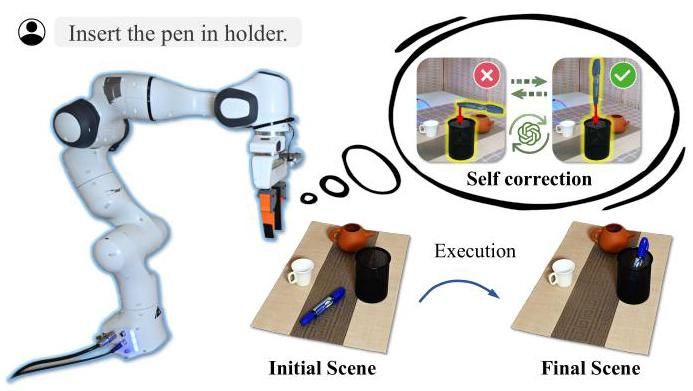


Figure 7. Closed-planning. Self-correction mechanism via RRC.

图7. 闭环规划。通过RRC的自我校正机制。

|  |  |
| --- | --- |
| Task | Success Rate |
| Pick up the cup on the dish | 95.24% |
| Recycle the battery | 91.30% |
| Insert the pen in holder | 86.36% |

|  |  |
| --- | --- |
| 任务 | 成功率 |
| 拿起盘子上的杯子 | 95.24% |
| 回收电池 | 91.30% |
| 将笔插入笔筒 | 86.36% |

Table 4. Behavior cloning with demonstrations from OmniManip.

表4. 使用OmniManip的演示进行行为克隆。

Closed-Loop Execution. Even with perfect planning, open-loop execution can still lead to task failure. Figure 8 illustrates two typical examples where planning succeeds, but open-loop execution causes failure. In the left image of Figure 8, the relative pose between the gripper and the object changes during the interaction, while the right image of Figure 8 shows a scenario where the target pose is dynamic, such as when the object moves during the task. To address these challenges, OmniManip employs pose tracking to enable real-time closed-loop execution. Recent work, ReKep, uses point tracking for closed-loop control but suffers from occlusions, leading to a failure rate [15]. In contrast, OmniManip demonstrates greater robustness to occlusions caused by object movement. This is a benefit of object-centric pose tracking, enabling continued tracking of canonical space interaction primitives based on the object pose, even when the primitives are no longer visible.

闭环执行。即使规划完美，开环执行仍可能导致任务失败。图8展示了两个典型例子，其中规划成功，但开环执行导致失败。在图8的左图中，夹爪与物体之间的相对姿态在交互过程中发生变化，而图8的右图展示了目标姿态是动态的场景，例如物体在任务过程中移动。为了解决这些挑战，OmniManip采用姿态跟踪实现实时闭环执行。最近的工作ReKep使用点跟踪进行闭环控制，但存在遮挡问题，导致 失败率[15]。相比之下，OmniManip对物体移动引起的遮挡表现出更强的鲁棒性。这是以物体为中心的姿态跟踪的优势，即使交互原语不再可见，也能基于物体姿态继续跟踪规范空间中的交互原语。

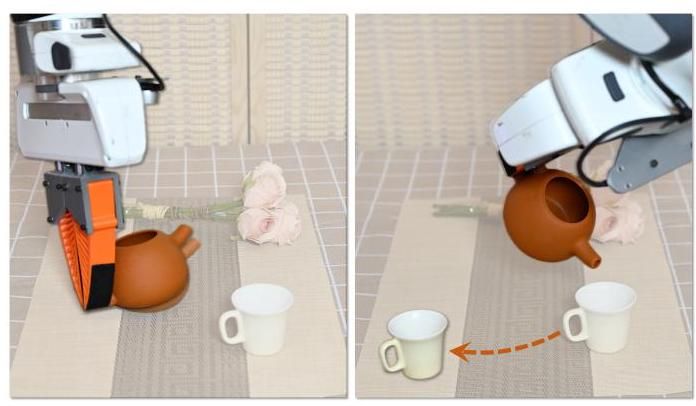


Figure 8. Two typical failure cases without closed-loop execution.

图8. 没有闭环执行的两个典型失败案例。

# 4.4. OmniManip for Demonstration Generation

# 4.4. OmniManip用于演示生成

We employed OmniManip to generate automatic demonstration data. Unlike prior methods reliant on task-specific privileged information, OmniManip collects demonstration trajectories for new tasks in a zero-shot manner, without needing task-specific details or prior object knowledge. To validate the effectiveness of OmniManip-generated data, we collected 150 trajectories per task to train behavior cloning policies [5]. These policies achieved high success rates, as shown in Table 4. Additional tasks and detailed results are provided in the appendix.

我们使用OmniManip生成自动演示数据。与之前依赖任务特定特权信息的方法不同，OmniManip以零样本方式为新任务收集演示轨迹，无需任务特定细节或先验物体知识。为了验证OmniManip生成数据的有效性，我们为每个任务收集了150条轨迹来训练行为克隆策略[5]。这些策略取得了较高的成功率，如表4所示。附录中提供了更多任务和详细结果。

# 5. Conclusion

# 5. 结论

In this work, we presented a novel object-centric intermediate representation that effectively bridges the gap between VLM and the precise spatial reasoning required for robotic manipulation. We structured interaction primitives in object canonical space to translate high-level semantic reasoning into actionable spatial constraints. The proposed dual closed-loop system ensures robust decision-making and execution, all without VLM fine-tuning. Our approach demonstrates strong zero-shot generalization across a variety of manipulation tasks, highlighting its potential for automating robotic data generation and improving the efficiency of robotic systems in unstructured environments. This work provides a promising foundation for future research into scalable, open-vocabulary robotic manipulation. Limitations. While advantageous, OmniManip also has limitations. It cannot model deformable objects due to pose representation. Its effectiveness also hinges on the mesh quality of 3D AIGC, which remains challenging despite progress. Additionally, multiple VLM calls present computational challenges, even with parallel processing.

在这项工作中，我们提出了一种新颖的以物体为中心的中间表示，有效弥合了VLM与机器人操作所需的精确空间推理之间的差距。我们在物体规范空间中构建了交互原语，将高级语义推理转化为可操作的 空间约束。所提出的双闭环系统确保了鲁棒的决策和执行，且无需VLM微调。我们的方法在各种操作任务中展示了强大的零样本泛化能力，突显了其在自动化机器人数据生成和提高非结构化环境中机器人系统效率方面的潜力。这项工作为未来可扩展、开放词汇的机器人操作研究提供了有前景的基础。局限性。尽管有优势，OmniManip也有局限性。由于姿态表示，它无法建模可变形物体。其有效性还取决于3D AIGC的网格质量，尽管有进展，这仍然具有挑战性。此外，即使采用并行处理，多次VLM调用也带来了计算挑战。

# Acknowledgments

# 致谢

We would like to thank Mingdong Wu and Tianhao Wu from PKU for their fruitful discussions, and Baifeng Xie from AgiBot for valuable technical support.

我们要感谢北京大学的吴明东和吴天昊的富有成果的讨论，以及AgiBot的谢柏峰提供的宝贵技术支持。

# References

# 参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.1, 2, 3

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, 等. Gpt-4技术报告. arXiv预印本 arXiv:2303.08774, 2023.1, 2, 3

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakr-ishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022. 2

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakr-ishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, 等. Rt-1: 用于大规模现实世界控制的机器人变压器. arXiv预印本 arXiv:2212.06817, 2022. 2

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023. 2

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, 等. Rt-2: 将网络知识转移到机器人控制的视觉-语言-动作模型. arXiv预印本 arXiv:2307.15818, 2023. 2

[4] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14455-14465, 2024.1,2

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, 和 Fei Xia. Spatialvlm: 赋予视觉-语言模型空间推理能力. 在IEEE/CVF计算机视觉与模式识别会议论文集, 页码 14455-14465, 2024.1,2

[5] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research, page 02783649241273668, 2023. 8

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, 和 Shuran Song. 扩散策略: 通过动作扩散进行视觉运动策略学习. 国际机器人研究杂志, 页码 02783649241273668, 2023. 8

[6] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shu-ran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. arXiv preprint arXiv:2402.10329, 2024. 6

程驰, 徐振佳, 潘楚尔, 埃里克·库辛, 本杰明·伯奇菲尔, 冯思远, 拉斯·特德拉克, 和宋书然. 通用操作界面: 无需野外机器人的野外机器人教学. arXiv预印本 arXiv:2402.10329, 2024. 6

[7] Qiyu Dai, Jiyao Zhang, Qiwei Li, Tianhao Wu, Hao Dong, Ziyuan Liu, Ping Tan, and He Wang. Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In European Conference on Computer Vision, pages 374-391. Springer, 2022. 2

戴启宇, 张佳瑶, 李启伟, 吴天昊, 董浩, 刘子源, 谭平, 和王鹤. 领域随机化增强的深度模拟与恢复用于感知和抓取镜面和透明物体. 欧洲计算机视觉会议, 第374-391页. Springer, 2022. 2

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023. 2

丹尼·德里斯, 夏飞, 梅赫迪·SM·萨贾迪, 科里·林奇, 阿坎莎·乔杜里, 布莱恩·伊克特, 阿伊赞·瓦希德, 乔纳森·汤普森, 范光, 余天和, 等. Palm-e: 一个具身多模态语言模型. arXiv预印本 arXiv:2303.03378, 2023. 2

[9] Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. arXiv preprint arXiv:2406.18915, 2024. 2

段佳飞, 袁文涛, 威尔伯特·普马凯, 王依茹, 基亚娜·埃萨尼, 迪特尔·福克斯, 和兰杰·克里希纳. 操纵一切: 使用视觉语言模型自动化现实世界机器人. arXiv预印本 arXiv:2406.18915, 2024. 2

[10] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu

方浩舒, 王晨曦, 方宏杰, 苟明浩, 刘继荣, 严恒旭, 刘文海, 谢一辰, 和卢策武

Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. IEEE Transactions on Robotics,

卢. Anygrasp: 在空间和时间域中稳健且高效的抓取感知. IEEE机器人学汇刊,

2023. 6

[11] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. The International Journal of Robotics Research, page 02783649241281508, 2023. 2

罗亚·菲鲁兹, 乔纳森·塔克, 斯蒂芬·田, 阿尼鲁达·马宗达, 孙建凯, 刘伟宇, 朱玉科, 宋书然, 阿什什·卡普尔, 卡罗尔·豪斯曼, 等. 机器人中的基础模型: 应用、挑战和未来. 国际机器人研究杂志, 第02783649241281508页, 2023. 2

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the world into large language models. Advances in Neural Information Processing Systems, 36:20482-20494, 2023. 2

洪一宁, 甄浩宇, 陈培豪, 郑树鸿, 杜一伦, 陈振芳, 和甘闯. 3d-llm: 将 世界注入大型语言模型. 神经信息处理系统进展, 第36卷:20482-20494, 2023. 2

[13] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. arXiv preprint arXiv:2403.08248, 2024. 2, 6

黄浩旭, 林凡奇, 胡英东, 王胜杰, 和高阳. Copa: 通过基础模型对零件空间约束的通用机器人操作. arXiv预印本 arXiv:2403.08248, 2024. 2, 6

[14] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973, 2023. 1, 6

黄文龙, 王晨, 张若涵, 李云竹, 吴佳俊, 和李飞飞. Voxposer: 用于语言模型机器人操作的可组合3D价值图. arXiv预印本 arXiv:2307.05973, 2023. 1, 6

[15] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. arXiv preprint arXiv:2409.01652, 2024. 2, 3, 6, 8

黄文龙, 王晨, 李云竹, 张若涵, 和李飞飞. Rekep: 用于机器人操作的关系关键点约束的时空推理. arXiv预印本 arXiv:2409.01652, 2024. 2, 3, 6, 8

[16] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical task and motion planning in the now. In 2011 IEEE International Conference on Robotics and Automation, pages 1470-1477. IEEE, 2011. 3

莱斯利·帕克·凯布林和托马斯·洛萨诺-佩雷斯. 当前的分层任务和运动规划. 2011年IEEE国际机器人与自动化会议, 第1470-1477页. IEEE, 2011. 3

[17] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Integrated task and motion planning in belief space. The International Journal of Robotics Research, 32(9-10):1194-1227, 2013. 3

莱斯利·帕克·凯布林和托马斯·洛萨诺-佩雷斯. 在信念空间中集成的任务和运动规划. 国际机器人研究杂志, 第32卷(9-10):1194-1227, 2013. 3

[18] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gam-bardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-world robot applications of foundation models: A review. Advanced Robotics, pages 1-23, 2024. 2

川原塚健人, 松岛达也, 安德鲁·甘巴德拉, 郭嘉贤, 克里斯·帕克斯顿, 和安迪·曾. 基础模型在现实世界机器人应用中的综述. 先进机器人学, 第1-23页, 2024. 2

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024. 2

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, 等. Openvla: 一个开源的视觉-语言-动作模型. arXiv 预印本 arXiv:2406.09246, 2024. 2

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015-4026, 2023. 3

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, 等. 分割一切. 在 IEEE/CVF 国际计算机视觉会议论文集, 第 4015-4026 页, 2023. 3

[21] Olivia Y Lee, Annie Xie, Kuan Fang, Karl Pertsch, and Chelsea Finn. Affordance-guided reinforcement learning via visual prompting. arXiv preprint arXiv:2407.10341, 2024. 2

Olivia Y Lee, Annie Xie, Kuan Fang, Karl Pertsch, 和 Chelsea Finn. 通过视觉提示进行可供性引导的强化学习. arXiv 预印本 arXiv:2407.10341, 2024. 2

[22] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. arXiv preprint arXiv:2402.12058, 2024. 4

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, 和 Yang Liu. 通过脚手架坐标促进大型多模态模型中的视觉-语言协调. arXiv 预印本 arXiv:2402.12058, 2024. 4

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In -

Junnan Li, Dongxu Li, Silvio Savarese, 和 Steven Hoi. Blip-2: 使用冻结图像编码器和大型语言模型进行语言-图像预训练. 在 -

ternational conference on machine learning, pages 19730- 19742. PMLR, 2023. 2

国际机器学习会议, 第 19730-19742 页. PMLR, 2023. 2

[24] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378, 2023. 2

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, 等. 视觉-语言基础模型作为有效的机器人模仿者. arXiv 预印本 arXiv:2311.01378, 2023. 2

[25] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023. 1

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, 等. Sphinx: 多模态大型语言模型的权重、任务和视觉嵌入的联合混合. arXiv 预印本 arXiv:2311.07575, 2023. 1

[26] Chang Liu, Kejian Shi, Kaichen Zhou, Haoxiao Wang, Jiyao Zhang, and Hao Dong. Rgbgrasp: Image-based object grasping by capturing multiple views during robot arm movement with neural radiance fields. IEEE Robotics and Automation Letters, 2024. 1

Chang Liu, Kejian Shi, Kaichen Zhou, Haoxiao Wang, Jiyao Zhang, 和 Hao Dong. Rgbgrasp: 通过捕捉机器人手臂运动期间的多个视图进行基于图像的物体抓取，使用神经辐射场. IEEE 机器人与自动化快报, 2024. 1

[27] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024. 2, 3

Fangchen Liu, Kuan Fang, Pieter Abbeel, 和 Sergey Levine. Moka: 通过基于标记的视觉提示进行开放词汇的机器人操作. 在 ICRA 2024 的第一届视觉-语言模型导航与操作研讨会, 2024. 2, 3

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 1, 2

Haotian Liu, Chunyuan Li, Qingyang Wu, 和 Yong Jae Lee. 视觉指令调优. 神经信息处理系统进展, 36, 2024. 1, 2

[29] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Ji-ayuan Gu, and Hao Su. One-2-3-45++: Fast single image to objects with consistent multi-view generation and 3d diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10072- 10083, 2024. 2, 4

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Ji-ayuan Gu, 和 Hao Su. One-2-3-45++: 快速从单图像生成 物体，具有一致的多视图生成和 3D 扩散. 在 IEEE/CVF 计算机视觉与模式识别会议论文集, 第 10072-10083 页, 2024. 2, 4

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023. 3

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, 等. Grounding dino: 将 DINO 与接地预训练结合用于开放集物体检测. arXiv 预印本 arXiv:2303.05499, 2023. 3

[31] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. arXiv preprint arXiv:2306.15724, 2023. 1

Zeyi Liu, Arpit Bahety, 和 Shuran Song. Reflect: 总结机器人经验以进行故障解释和纠正. arXiv 预印本 arXiv:2306.15724, 2023. 1

[32] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In The International Symposium of Robotics Research, pages 132-157. Springer, 2019. 2

Lucas Manuelli, Wei Gao, Peter Florence, 和 Russ Tedrake. kpam: 类别级机器人操作的关键点功能. 在国际机器人研究研讨会, 第132-157页. Springer, 2019. 2

[33] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. arXiv preprint arXiv:2402.07872, 2024. 2, 3

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, 等. Pivot: 迭代视觉提示引出VLM的可操作知识. arXiv预印本 arXiv:2402.07872, 2024. 2, 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748-8763. PMLR, 2021. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, 等. 从自然语言监督中学习可迁移的视觉模型. 在国际机器学习会议, 第8748-8763页. PMLR, 2021. 1

[35] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspon-

Tanner Schmidt, Richard Newcombe, 和 Dieter Fox. 自监督视觉描述符学习用于密集对应.

dence. IEEE Robotics and Automation Letters, 2(2):420-

IEEE机器人与自动化快报, 2(2):420-

427, 2016. 2

[36] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)- equivariant object representations for manipulation. In 2022 International Conference on Robotics and Automation (ICRA), pages 6394-6400. IEEE, 2022. 2

Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, 和 Vincent Sitzmann. 神经描述符场: 用于操作的SE(3)等变对象表示. 在2022年国际机器人与自动化会议 (ICRA), 第6394-6400页. IEEE, 2022. 2

[37] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 11523-11530. IEEE, 2023. 1

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, 和 Animesh Garg. Progprompt: 使用大语言模型生成情境化机器人任务计划. 在2023年国际机器人与自动化会议 (ICRA), 第11523-11530页. IEEE, 2023. 1

[38] Ioan A Sucan, Mark Moll, and Lydia E Kavraki. The open motion planning library. IEEE Robotics & Automation Magazine, 19(4):72-82, 2012. 2

Ioan A Sucan, Mark Moll, 和 Lydia E Kavraki. 开放运动规划库. IEEE机器人与自动化杂志, 19(4):72-82, 2012. 2

[39] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, et al. Curobo: Parallelized collision-free robot motion generation. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 8112-8119. IEEE, 2023. 2

Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, 等. Curobo: 并行化无碰撞机器人运动生成. 在2023年国际机器人与自动化会议 (ICRA), 第8112-8119页. IEEE, 2023. 2

[40] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. arXiv preprint arXiv:2403.02151, 2024. 2, 4

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, 和 Yan-Pei Cao. Triposr: 从单张图像快速重建3D对象. arXiv预印本 arXiv:2403.02151, 2024. 2, 4

[41] Marc Toussaint, Jason Harris, Jung-Su Ha, Danny Driess, and Wolfgang Hönig. Sequence-of-constraints mpc: Reactive timing-optimal control of sequential manipulation. In 2022 IEEE/RSJ Ineternational Conference on Intelligent Robots and Systems (IROS), pages 13753-13760. IEEE, 2022. 2

Marc Toussaint, Jason Harris, Jung-Su Ha, Danny Driess, 和 Wolfgang Hönig. 序列约束MPC: 顺序操作的响应时间最优控制. 在2022年IEEE/RSJ国际智能机器人与系统会议 (IROS), 第13753-13760页. IEEE, 2022. 2

[42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 1

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, 等. Llama 2: 开放基础和微调的聊天模型. arXiv预印本 arXiv:2307.09288, 2023. 1

[43] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15964-15973, 2021. 6

Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, 和 Cewu Lu. 在杂乱中发现抓取性以实现快速准确的抓取检测. 在IEEE/CVF国际计算机视觉会议, 第15964-15973页, 2021. 6

[44] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. arXiv preprint arXiv:2201.12716, 2022. 3

Bowen Wen, Wenzhao Lian, Kostas Bekris, 和 Stefan Schaal. 你只需演示一次: 从单次视觉演示进行类别级操作. arXiv预印本 arXiv:2201.12716, 2022. 3

[45] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified pose estimation and tracking of novel objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17868- 17879, 2024. 3

Bowen Wen, Wei Yang, Jan Kautz, 和 Stan Birchfield。Foundationpose:统一 姿态估计与新物体跟踪。在IEEE/CVF计算机视觉与模式识别会议论文集，第17868-17879页，2024年。3

[46] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. arXiv preprint arXiv:2401.00025, 2023. 2

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, 和 Pieter Abbeel。任意点轨迹建模用于策略学习。arXiv预印本arXiv:2401.00025，2023年。2

[47] Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, and Hao Dong. Canonical representation and force-based pretraining of tactile for dexterous visuo-tactile policy learning. arXiv preprint arXiv:2409.17549, 2024. 1

Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, 和 Hao Dong。规范表示与基于力的 触觉预训练用于灵巧视觉触觉策略学习。arXiv预印本arXiv:2409.17549，2024年。1

[48] Tianhao Wu, Mingdong Wu, Jiyao Zhang, Yunchong Gan, and Hao Dong. Learning score-based grasping primitive for human-assisting dexterous grasping. Advances in Neural Information Processing Systems, 36, 2024. 1

Tianhao Wu, Mingdong Wu, Jiyao Zhang, Yunchong Gan, 和 Hao Dong。学习基于分数的抓取原语用于辅助人类灵巧抓取。神经信息处理系统进展，36，2024年。1

[49] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441, 2023. 3

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, 和 Jianfeng Gao。标记集提示释放GPT-4V中的非凡视觉定位能力。arXiv预印本arXiv:2310.11441，2023年。3

[50] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for lidar understanding. arXiv preprint arXiv:2312.14074, 2023. 2

Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, 和 Shanghang Zhang。Lidar-llm:探索大语言模型在 激光雷达理解中的潜力。arXiv预印本arXiv:2312.14074，2023年。2

[51] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. arXiv preprint arXiv:2303.04129, 2023. 2

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, 和 Dale Schuurmans。决策基础模型:问题、方法与机遇。arXiv预印本arXiv:2303.04129，2023年。2

[52] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousa-vian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721, 2024. 2

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, 和 Dieter Fox。Robopoint:用于机器人空间功能预测的视觉语言模型。arXiv预印本arXiv:2406.10721，2024年。2

[53] Yiming Zeng, Mingdong Wu, Long Yang, Jiyao Zhang, Hao Ding, Hui Cheng, and Hao Dong. Lvdiffusor: Distilling functional rearrangement priors from large models into diffusor. IEEE Robotics and Automation Letters, 2024. 1

Yiming Zeng, Mingdong Wu, Long Yang, Jiyao Zhang, Hao Ding, Hui Cheng, 和 Hao Dong。Lvdiffusor:从大模型中蒸馏功能重排先验到扩散器。IEEE机器人与自动化快报，2024年。1

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975-11986, 2023. 1

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, 和 Lucas Beyer。语言图像预训练的Sigmoid损失。在IEEE/CVF国际计算机视觉会议论文集，第11975-11986页，2023年。1

[55] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. Advances in Neural Information Processing Systems, 36, 2024. 2

Jiyao Zhang, Mingdong Wu, 和 Hao Dong。通过扩散模型生成类别级物体姿态估计。神经信息处理系统进展，36，2024年。2

[56] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In European Conference on Computer Vision, pages 199-216. Springer, 2025. 2, 4, 6

Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, 和 Hao Dong。Omni6dpose:通用6D物体姿态估计与跟踪的基准与模型。在欧洲计算机视觉会议，第199-216页。Springer，2025年。2, 4, 6

[57] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10324-10335, 2024. 4

Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, 和 Song-Hai Zhang。三平面遇见高斯溅射:使用变压器进行快速且可泛化的单视图3D重建。在IEEE/CVF计算机视觉与模式识别会议论文集，第10324-10335页，2024年。4

1. \*: Equal contributions. : Corresponding author

   \*: 同等贡献。 : 通讯作者 [↑](#footnote-ref-29)