arXiv:2303.02871v1 [cs.RO] 6 Mar 2023

为视觉与语言操纵命名物体

Tokuhiro Nishikawa[1*]、Kazumi Aoyama[1*]、Shunichi Sekiguchi[1*]、Takayoshi Takayanagi[1]、Jianing Wu[1]、Yu Ishihara[1]、、Tamaki Kojima[1]和 Jerry Jun Yokono[1]

摘要——基于自然语言指令的机器人操作任务需要人类与机器人对目标物体形成共同理解。然而，由于指令缺少重要信息，或未能正确表达完成任务所需的目标物体，这些指令往往存在解释歧义。为了解决这一歧义问题，我们提出假设：提前为目标物体“命名”可以减少自然语言指令的歧义。我们提出一种机器人系统和方法，提前将名称与物体的外观相结合，从而在后续操作任务中，可以使用物体的唯一名称执行指令，以便轻松消除物体歧义。为验证该方法的有效性，我们构建了一个能够记忆目标物体的系统，并展示了为物体命名有助于检测目标物体，同时能够提高操作指令的成功率。在有歧义的指令中，采用该方法后，物体操作任务的成功率提高了31%。

I. 引言

对于能够执行多种任务的家庭服务机器人而言，自然语言接口非常重要，因为这是人类无需特殊训练即可自由交流的最简单方式之一。因此，基于自然语言接口的机器人操作得到了广泛研究[1], [2], [3], [4], [5]。

抓取与放置等操作是家庭服务机器人最重要的任务。为了通过自然语言指令执行这些任务，人类与机器人之间需要有共同的理解。然而，自然语言指令往往缺少重要信息，或者未能正确表达目标物体，因而无法完成任务。

我们的研究发现，17% 的人工编写的物体操作指令在目标物体的表达上存在歧义，尽管我们要求标注人员尽可能详细地描述物体操作指令，以便机器人能够识别目标物体（详情见第 IV 节）。本文将操作指令歧义定义如下：1）目标物体存在多个候选对象。2）目标物体的指代表达有误。示例见图1。此外，Hatori 等人 [5] 报告称，21% 的物体操作指令被判定存在多个目标物体候选对象，因此被视为含糊不清的指令。因此，使用自然语言的物体操作指令容易包含歧义表达，导致机器人无法识别目标物体，

*这些作者对本研究的贡献相同。[1]索尼集团公司，日本东京港区港南 1-7-1，邮编108-0075。通讯联系：tokuhiro.nishikawa@sony.com

指令：“把瓶子拿起来。”　　指令：“把猫抱起来。”



图1. 包含有歧义表达的自然语言指令示例：**左：**目标物体存在多个候选对象的指令。（视野中有三个瓶子。）**右：**目标物体的表达不正确。（视野中没有猫，但有一只狗。）

从而无法按预期执行操作指令。

为解决这些歧义问题，已有多种方法提出让机器人向操作员提供额外反馈 [5], [6], [7]。然而，这些方法每次遇到歧义时都需要额外的交互步骤来明确指令，而我们认为可以通过对其进行命名来省略这些步骤。

因此，我们提出假设：提前对目标物体进行“命名”将减少自然语言指令的歧义。一旦机器人和人类共享物体的名称，就可以通过该名称唯一确定目标物体。因此，与先前的方法不同，无需额外交互 [5], [6], [7]。我们提出了一种融入物体外观与名称的机器人系统和方法，使得在后续操作任务中，可以使用目标物体的唯一名称执行指令，从而轻松消除物体之间的歧义。为证明该方法的有效性，我们构建了一个能够记忆目标物体外观和名称的机器人系统，并展示了对物体进行命名有助于检测目标物体，并提高操作指令的成功率。通过该方法，在有歧义的指令下，物体操纵任务的成功率提高了31%。

总而言之，本文的贡献如下：

- 提出一种方法：预先根据物体的外观为其命名，有助于减少自然语言物体操作指令中目标物体的歧义。
- 提出一种系统，该系统能够记忆物体的名称及其外观，并对其进行回忆。
- 通过仿真和真实机器人系统中的实验，证明上述提案在机器人领域是有效的

自然语言操作任务。

II. 相关工作

已有许多研究致力于使机器人能够遵循自然语言指令 [1], [2], [3],[4], [5], [7], [8], [9]。自然语言指令操作是该领域的热门研究课题之一。例如，Hatori 等人 [5] 构建了一个交互系统，用户可以通过无约束口语指令来执行物体拾取任务。CLIPort [4] 能够解决多种由语言指令驱动的任务，从打包物体到折叠衣物。

指代表达理解（REC） [10] 与自然语言指令机器人操作密切相关。REC旨在根据指代表达，在图像中定位目标物体。在视觉语言操作中，也有必要在指令中定位物体。MDETR [11] 和 GLIP [12]能够基于经过大规模数据训练的大模型，估计自然语言所指物体的位置。

在机器人操作及其指令视觉定位过程中，系统必须识别目标物体。然而，人类常常会含糊地指称目标物体，导致系统无法正确定位。当人类语言指令存在歧义时，机器人需要利用指令以外的信息消除歧义，从而确定目标物体。例如，当某人发出“把我的瓶子递给我”这一指令时，如果有多个瓶子可供选择，机器人可以通过多种方式确认“我的瓶子”是哪一个：反问此人、识别此人的注视或指向手势，或者参考过去记忆。一些研究提出通过基于语言的反馈来消除机器人指令中的歧义 [5], [7], [6]。Whitney等人 [6] 表明，当机器人无法确定某个物体时，简单地提出一个确认问题，例如“是这个吗？”，会有所帮助。Hatori 等人 [5] 表明，请求对方改述指令并重新进行解释，可以提高对所指物体的识别性能。Yang等人 [7]通过开发一种能够借助对话消除歧义的抓取系统，证明使用物体属性有助于消除歧义。然而，尽管语言反馈是消除歧义的有效工具，但针对每条有歧义的指令进行额外交互，可能会降低机器人的易用性。

人类在日常交流中处理歧义的一种方式是词汇协调。词汇协调是一种心理现象，指人们倾向于采用交谈对象所使用的术语。Iio等人发现，人机交互中也会出现词汇协调[13]。例如，当机器人用物体的颜色来指代物体时，人类也会用物体的颜色来指代它；当机器人用物体的名称来称呼它时，人类也会用物体的名称来称呼它。他们认为，促进人类与机器人之间的词汇协调，即用共同术语称呼物体，可以提高机器人对所指物体的识别性能，但他们并未开展实验来验证这一假设。

至于提出物体命名概念的相关研究，Nakamura 等人 [14]指出，物体有些名称仅在家庭内部使用；他们认为，人类和机器人能够用相同的名称理解和指称物体是理想的。他们提出了一种方法，可从视听输入中学习新物体及其名称，并采用 RoboCup@Home 联赛 [15]中的一项任务对该方法进行了评估。然而，他们并未明确展示使用名称对机器人任务的优势。Jiang 等人 [3]训练了一个大规模视觉与语言模型，并展示了通过结合图像和语言的指令，可以抓取已命名的物体。然而，他们并未展示使用名称的益处。目前已有一些关于语言指令操作的基准测试[16], [17], [18]。然而，这些基准测试并不侧重于解决人类语言歧义。

因此，目前尚无文献展示人类与机器人共享物体名称的益处。本文提出了一种新颖的系统，该系统能够记住物体的名称，并展示了为物体命名有助于检测目标物体，并提高操作指令的成功率。

III. 为视觉与语言操纵命名物体

我们的研究目标是构建一个能够根据用户提供的自然语言指令执行抓取与放置的机器人系统。我们的系统按照以下步骤运行。

1）接收自然语言指令（文本）和场景信息（RGB-D 图像）。2）预测要抓取的物体和放置位置。3）利用预测信息操作物体。

然而，自然语言指令的歧义使得难以在步骤 2）中识别目标物体。如果系统未能预测要抓取的物体和放置位置，就无法按照用户的意图运行。因此，为了减少此类歧义并提升物体操作性能，我们提出预先为物体命名，并在指令中使用该名称，以正确识别抓取物体和放置位置。通过预先为物体命名，可以缓解语言指令的歧义，使系统能够完成拾取与放置任务。所提出的算法由两个主要过程组成。

- **命名过程**
  1) 接收命名指令和场景图像。2) 检测场景中的待命名物体，以及指令中对应的名称。3) 将该物体的图像及其名称存入记忆。

- **操作过程**
  1）接收一条抓取与放置指令以及一幅场景图像。
  2）找到要抓取的物体和放置位置。

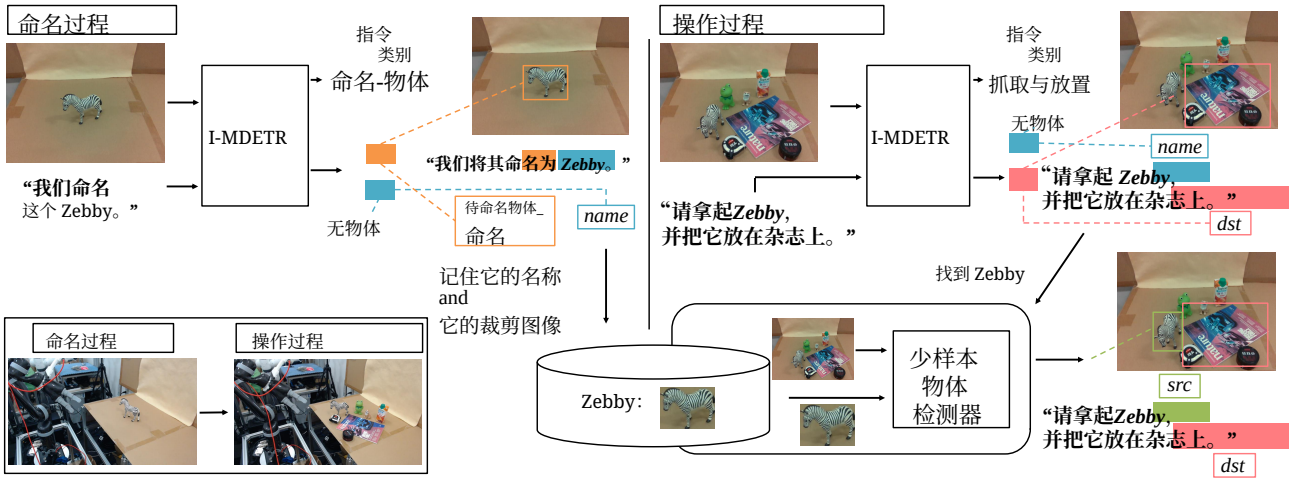


图2. 我们提出了一种系统，可预先为目标物体命名，并在抓取与放置指令中使用该名称。在命名过程中，如左侧所示，我们首先使用 I-MDETR 检测场景中待命名的物体及指令中的物体名称，然后将其图像和名称存储在记忆中。命名过程结束后，系统会接收包含该名称的抓取与放置指令，并如右侧所示，使用 I-MDETR 和少样本目标检测器找到要抓取的物体及其放置位置。

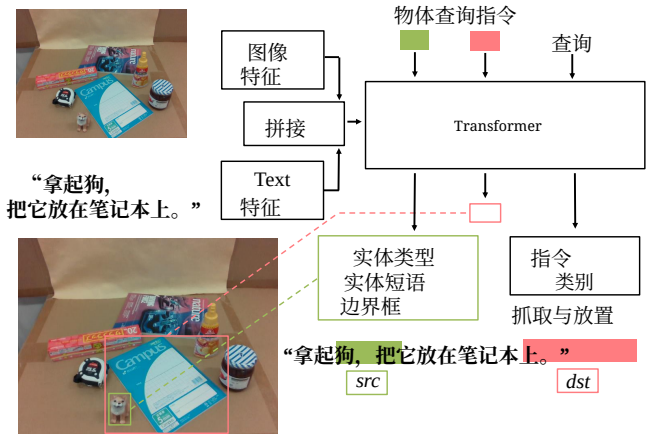


图3. 指令-MDETR（I-MDETR）概览。I-MDETR 是 MDETR [11] 的扩展版本。它会预测指令类别、实体类型及其对应的短语，以及物体的边界框。

3）根据在 2）中提取的信息操控物体。

为实现命名和操作过程中的步骤 2），我们引入了指令 MDETR（I-MDETR），这是一种基于 MDETR [11]、针对语言指令扩展的物体检测器。此外，为了实现操作过程中的步骤 2），我们还实现了一种少样本目标检测器[19], [20]，利用记忆图像，从 I-MDETR 检测到的物体中找到指令中命名的物体。在以下小节中，我们将详细介绍 I-MDETR，并解释命名和操作过程的算法流程。

A. 指令 *MDETR*（*I-MDETR*）

I-MDETR 是 MDETR [11]，的扩展。MDETR 是一种为指代表达理解（REC）任务提出的方法，它可以预测图像中物体的边界框，并将其与文本短语进行视觉定位。与 MDETR 相比，I-MDETR 还会预测实体类型，用以表示该短语所扮演的角色在指令中。图3展示了 I-MDETR 的架构。我们为当前任务设置定义了以下实体类型。

- **名称**：物体的名称。
- **待命名物体**：要命名的物体。
- **源物体**：抓取的目标物体。
- **目标**：指示放置位置的目标物体。

请注意，可以添加其他实体类型，以涵盖不同的任务设置。此外，为了对指令本身的目标进行分类，I-MDETR 还会预测指令类别。我们为任务定义了以下指令类别。

- **命名物体**：用于命名的指令。
- **抓取与放置**：用于抓取与放置的指令。
- **不支持的指令**：不支持的指令。

不支持的指令 被引入，以妥善处理与我们的操作任务无关的指令。请注意，除实体类型外，还可以添加其他指令类别，以覆盖不同的任务设置。

B. 命名过程

当 I-MDETR 将**命名物体** 预测为指令类别时，我们首先从 I-MDETR 的输出中提取指令里的实体类型**名称** 和**待命名物体** 短语。具体而言，我们会搜索包含**待命名物体** 短语的边界框，并检查其实体类型，以确定边界框中物体的名称（另见图2）。如果系统未能找到目标物体及其名称，就会在此处结束处理请求。否则，当系统找到物体及其名称后，会将边界框中的物体保存为图像，并以指令中提供的名称为图像添加标签。通过用物体名称标记图像，在操作过程中，我们可以使用少样本检测器将边界框中的物体与图像进行匹配，从而找到已命名物体。

C. 操作过程

当 I-MDETR 将**抓取与放置** 预测为指令类别时，系统会按照以下流程执行操作。首先，系统利用 I-MDETR 检测到的边界框和短语（及其实体类型），搜索要抓取的目标物体和目标放置位置。通过在指令中搜索实体类型**源物体** 或**名称**，来选择要抓取的目标物体。如果在指令中找到**源物体**，则将图像中对应物体的边界框作为要抓取的目标物体。然而，如果在指令中找到实体类型**名称**，I-MDETR 输出的边界框可能不正确。这是因为 I-MDETR 的训练数据中没有物体名称。因此，我们使用少样本目标检测器[19], [20]，在指令中查找具有指定名称的物体（有关少样本目标检测器的详细信息，请参见附录 II）。具体而言，我们使用少样本检测器在场景中查找与已保存图像相匹配的物体，并将匹配度最高的物体作为目标物体。如果在指令中找到多个实体类型**源物体** 或**名称**，我们会根据 I-MDETR 输出的置信度分数，选择概率最高的实体类型。目标放置位置的选择方式类似，即根据指令中的**目标** 或**名称** 进行选择。确定目标物体和放置位置后，我们将机械臂预定义的操作技能集结合起来，执行抓取与放置操作。

*D. I-MDETR*及其他组件的训练

在训练 I-MDETR 时，我们使用物理引擎 MuJoCo [21] 实现了自动生成训练数据集的流程。我们对 60 种常见家居物体进行了三维扫描，并将它们摆放在 MuJoCo仿真环境中的桌子上。脚本基于人类标注的表达模板，自动生成指令文本和标注。我们基于使用 RefCOCOg [22]，预训练的 MDETR 权重，使用生成的数据集对 I-MDETR 模型进行了微调。RefCOCOg 是视觉与语言领域中广泛用于 REC 的数据集。I-MDETR 训练数据集的详细信息见附录 I。我们还在附录 III 中附上了抓取与放置方法的详细信息。

IV. 实验评估

我们的实验旨在回答以下问题：提前根据物体的外观为其命名，是否有助于减少自然语言指令中目标物体的歧义，并提高操作的成功率?

我们的实验流程如图4所示。我们比较了无命名和含命名情况下的操作表现。

- **无命名**：只需指示机器人执行抓取与放置任务，并向机器人系统提供场景图像和指令文本。
- **含命名**：首先指示机器人记住一个物体及其名称，然后指示机器人根据该名称执行抓取与放置任务。

在含命名的设置中，用户命名物体时，只向机器人展示一个物体，并告诉它这个物体的名称。在命名设置中，

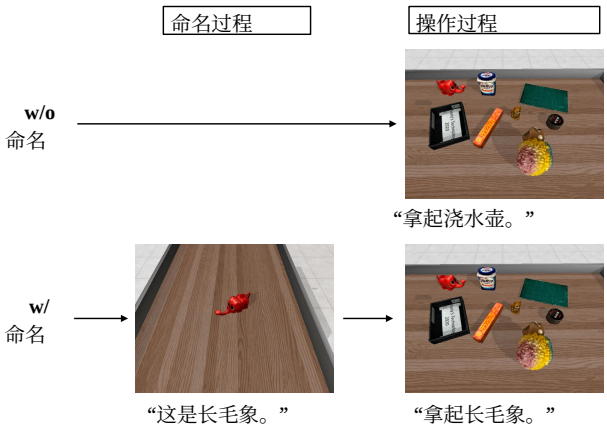


图4. 实验流程。在无命名条件（**无命名**）下，我们只需指示机器人执行抓取与放置任务。在含命名条件（**含命名**）下，我们先告诉机器人记住物体及其名称，然后指示它根据该名称执行抓取与放置任务。

I-MDETR 识别出物体（待命名物体）后，会将该物体的裁剪图像存储在记忆中。在我们的实验中，我们假设机器人可以从不同视角多次观察该物体。实验中，记忆中保存了 4 张来自不同视角的图像。

我们在仿真和真实机器人环境中，使用人类生成的语言指令进行了实验。以下各节将详细介绍人工指令收集、评估指标和结果。

A. 人工指令收集

我们使用了与训练数据集相同的 60 个物体来生成评估场景。从 60 个物体中随机选取 6 到 8 个，并将它们放在桌子上。相机位置和姿态的选取方式与生成训练数据集时相同。总共生成了 20 个评估场景。

我们请 15 位志愿者为每个场景标注一条指令，共收集了 300 条指令（每个场景 15 条指令）。我们要求志愿者遵循以下说明。

- 用自然语言撰写抓取与放置指令。
- 请勿撰写需要抓取和/或放置多个物体的指令。
- 请勿撰写要求将物体移离桌子的指令，例如“拿起罐子并把它放进冰箱。”
- 目标物体应当能够被唯一识别。

以这种方式从人类志愿者处收集评估指令有两个原因。其一是研究人类给出有歧义的指令的频率，另一则是研究他们会给出何种有歧义的指令。以往研究表明，即使被要求唯一指代某个物体，人类有时仍会给出有歧义的指令 [5], [10], [22]。

收集完成后，我们请一位经验丰富的人工标注员检查全部300条指令，以确定其中有多少条在识别目标物体时存在歧义。结果发现，51条指令存在歧义，占总数的17%，

其中31个案例是指令无法唯一确定物体（即存在多个候选物体），另有24个案例是由于误解目标物体而未能准确指代物体。我们确认，即使要求志愿者标注抓取与放置指令时，确保操作员（即机器人）能够在不存在解释歧义的情况下唯一确定目标物体，第三方（即经验丰富的标注员）严格检查后，仍发现部分指令存在歧义。换言之，人类指令中的歧义很难消除，因此除非机器人能够消解这些歧义，否则就无法成功抓取和放置物体。

指令类别、边界框、实体短语和实体类型也均由标注员手动标注。即使指令中的目标物体有歧义，志愿者仍有意操控某个物体。因此，对于这些有歧义的指令，标注员会推测志愿者的意图并选定一个目标。

尽管我们原本预计志愿者会编写拿取指令，但收集到的部分指令实际上是我们的机器人无法执行的，例如“把笔记本平衡地放在瓶子上”。这类指令的指令类别会被手动替换为“**不支持的指令**”，因为我们认为机器人需要能够忽略这类指令，尤其是在实际应用中。通过上述流程创建的这些人工标注指令用于实验中的无命名条件。

基于无命名条件的评估数据，我们生成了实验中含命名条件的评估数据。我们只需将指令中的指代表达替换为名称。物体的名称由志愿者观察物体后手动标注。命名指令通过从预定义的命名句式模板中随机选取得到（例如：“那个玩具的名称是 A”或“我叫它 A”）。

B. 评估指标

我们定义了以下评估指标。

- **ICR**：指令分类成功率，即 I-MDETR 正确预测的指令类别数占指令总数的比例。
- **PR**：短语提取成功率，即 I-MDETR 从指令中正确提取目标物体短语的比例。根据物体的实体类型，该短语可以是物体名称或指代表达。
- **BR**：边界框检测成功率，即 I-MDETR 在输入图像中成功检测到目标物体区域的指令数占指令总数的比例。如果检测到的边界框与真实标注的交并比大于 0.5，则视为成功。
- **成功率**：流程成功率。系统成功完成给定流程（命名或操作）的指令数占指令总数的比例。在含命名的情况下，要成功完成操作流程，机器人必须先成功完成命名流程。

C. 仿真实验结果

如第 IV-A 节所述，共有 51 条含糊指令和 249 条明确指令，总计 300 条。我们使用这些指令分别进行了不含命名和含命名的实验。表 I 展示了结果。

从表 I 中我们可以确认，在含命名的情况下，全部 300 条指令的操作过程**SR** 为 69.7%，相比无命名时的 57.0% 提升了 12.7%。聚焦于 51 条含糊指令，含命名时操作过程的**SR** 为 62.7%，而无命名时为 31.4%。这几乎实现了 2 倍的性能提升。这些结果表明，对物体进行命名可以有效缓解指令的歧义。进一步详细分析结果可以发现，含命名和无命名时含糊指令的**PR** 相同，而 **BR** 提升了 35.2%，从而促成了**SR** 的提升。我们认为 **BR** 提升的原因在于命名过程降低了歧义。对于含糊指令，可以从指令文本中提取目标物体的实体短语，但由于存在歧义，边界框检测往往会失败。在无命名的情况下，I-MDETR 应根据给定的含糊指令检测物体。相反，在含命名的情况下，少样本目标检测器利用命名过程中预先记忆的物体图像来检测目标。这些结果表明，当指令中以含糊的方式指代目标物体时，仅凭指代表达很难找到目标物体。

此外，对于 249条明确指令，通过命名，**操作过程的成功率** 也提高了 8.9%。**提取率** 和 **边界框检测率** 分别提高了 14.0% 和 7.1%。**提取率** 的提升表明，提取名称比提取指代表达更容易。**提取率**的提升也带来了**边界框检测率** 的提高。这是因为，采用命名后，系统会根据提取出的名称检索目标物体记忆中的图像，并运行少样本物体检测器。因此，命名似乎不仅有助于消除歧义，也能让目标物体更容易被检测到。

图 5 展示了无命名时的定性结果，说明机器人无法消除人类指令中的歧义。图 6 展示了含命名时的定性结果。对比图 5 和图 6 可见，为物体命名有助于机器人解决人类指令中的歧义。

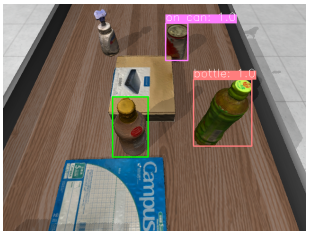
D. 真实机器人实验

我们使用 Kinova Jaco2 [23]机械臂和 Realsense D435i [24] 相机，在真实机器人环境中评估了我们的系统。我们使用仿真实验中用到的真实物体测试了该方法。评估流程如下：

1）从实际物体中抽取 1 个物体，并将抽取的物体放在桌子上。2）向系统输入场景的命名指令和 RGB-D 图像。

表 I 仿真实验结果

		命名过程		操作过程		
		成功率 [%]	ICR [%]	提取率 [%]	边界框检测率 [%]	成功率 [%]
无命名	所有指令	-	95.3 (286/300)	72.0 (216/300)	57.0 (171/300)	57.0 (171/249)
	无歧义	-	95.2 (237/249)	71.5 (178/249)	63.0 (157/249)	62.2 (155/249)
	有歧义	-	96.1 (49 / 51)	74.5 (38 / 51)	27.5 (14 / 51)	31.4 (16 / 51)
含命名	所有指令	98.3 (295/300)	94.0 (282/300)	83.7 (251/300)	69.3 (208/300)	69.7 (209/300)
	无歧义	98.8 (246/249)	94.4 (235/249)	85.5 (213/249)	70.1 (176/249)	71.1 (177/249)
	模糊不清	96.1 (49 / 51)	92.2 (47 / 51)	74.5 (38 / 51)	62.7 (32 / 51)	62.7 (32 / 51)



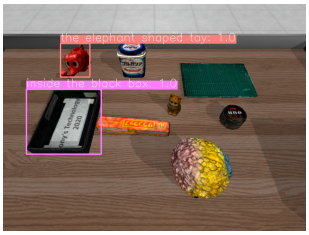
(a)**"拿起瓶子并放到罐子上。"** 有多个瓶子。目标是中心处的棕色瓶子。**存在歧义，失败**。



(b)**"拿起柴犬并把它放到果酱罐上。"**没有狗，只有一只猫，而且对"柴犬"的检测失败。**存在歧义，失败**。



(c)**"拿起彩色玩具并把它放到盒子上。"**这个词"彩色"似乎是主观的。机器人拿起了绿色青蛙玩具，但预期目标是另一个玩具。**无歧义，失败**。



(d)"拿起大象形状的玩具并将其放进黑色盒子里。" 大象形状的物体是一个浇水壶，而它是否属于玩具是一个主观判断。**无歧义，成功**。

图 5. **无命名** 示例。绿色边界框是真实标注。粉色和橙色边界框分别是预测的**源物体** 和**目标**。



(a)**"拿起*Kaki Shoyu* 并放在罐子上。"**原始指令是**"拿起瓶子并放在罐子上。"**而且有多个瓶子。得益于命名，不会产生误解在人类和机器人之间，这时间。**含糊不清，成功**。



(b)**"拿起*Maru chan* 并放到果酱罐上面。"**原始指令是**"拿起柴犬并把它放到上面的果酱罐上。"**而且没有柴犬。这一次，指令被机器人正确理解了。**存在歧义，成功**。



(c)**"拿起*RaimbowPup* 并把它放到盒子上。"**原始指令是**"拿起这只彩色玩具并将其放在盒子。"**机器人拿起了绿色青蛙玩具，拿错了。通过使用指令中的名称，它成功了。**无歧义，成功**。



(d)**"牙刷需要被内部 *Sara*。"** 短语" A 需要位于 B 内部"未被识别为抓取与放置指令。**无歧义，失败**。

图 6。 **w/ 命名**示例。绿色边界框为真实标注。粉色和橙色边界框是预测的**源物体** 和**d** **st**.

表II

真实机器人实验结果

成功率 [%]	
命名过程	操作过程
95.0（19/20）	45.0（9/20）

3）抽取 6 至 8 个物体（包括我们在 1）中选定的物体），并将它们放在桌子上。

4) 向系统输入一条包含 2) 中给出的物体名称的抓取与放置指令，以及该场景的 RGB-D 图像。

5) 检查机器人是否完成了指令。

我们请 1 名志愿者为每个场景提供指令，并将此流程运行了 20 次。

真实机器人实验的定量结果如表。II。图7展示了真实机器人的一个运动示例以及系统的预测。从表。II和图7中，我们可以确认我们的系统在真实机器人环境中也能运行。然而，成功率低于仿真中的成功率。原因之一是仿真到真实迁移过程中性能下降。I-MDETR 使用仿真器生成的数据集进行训练。提升系统在真实机器人环境中的性能是我们未来的工作。实际的机器人运动也可以在补充视频中观看。

V. 结论

为了检验我们的假设——提前为目标物体命名可以减少自然语言指令中的歧义，我们构建了一个能够记住

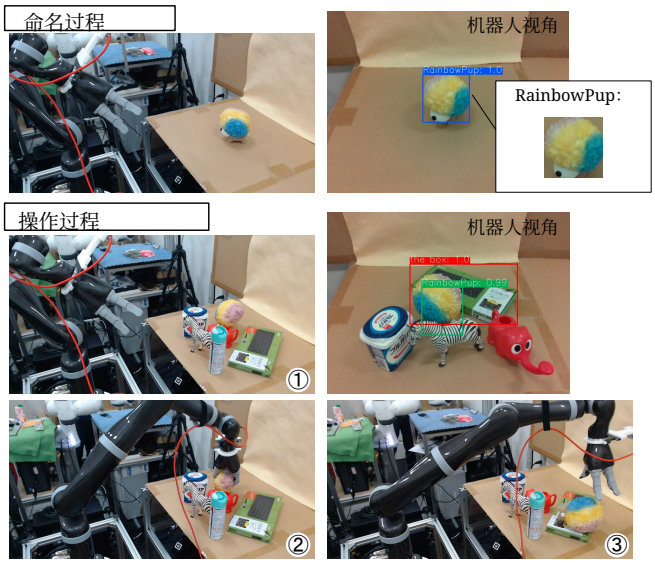


图7。真实机器人实验结果示例。命名指令为**“它的昵称是*RainbowPup*。”**，抓取并放置指令为**“拿起*RainbowPup* 并将其放在盒子上。”**。

物体名称及其外观，并能理解使用该名称的指令的机器人系统。我们的实验表明，为目标物体命名并使用名称对其进行操作，能够显著提高操作的成功率。命名不仅可以减少指令中的歧义，我们还认为，它也有助于通过自然语言给出复杂指令，例如运动轨迹控制。探究命名的其他积极作用是我们未来的研究方向。

附录 I I-MDETR 训练数据集生成

我们使用物理仿真器 MuJoCo [21] 为 I-MDETR 生成训练数据集。我们通过 iOS 应用 Metascan [25]扫描了 60 种常见家居物体。使用体积层次近似凸分解 [26]，构建碰撞体积后，我们将这些物体导入 MuJoCo [21]。除了扫描得到的物体外，桌子和地板等均基于 metaworld[27] 环境。Metaworld 是一个用于机器人研究中多任务学习的开源仿真基准。

我们的数据集生成分为两个步骤：场景生成，以及为每个场景进行指令标注。在本研究中，我们假设物体放置在桌子上，例如 [4]。为了生成场景，我们首先随机抽取 1 到 10 个物体并将其放在桌子上。我们还从预定义的 4 个相机位姿中随机抽取一个，这些位姿分别使相机从右侧、左侧、前侧和后侧朝向桌子，然后渲染场景。我们为 I-MDETR 训练数据集生成了 10,000 个场景。

参照每个场景中的物体信息，指令生成器会自动生成指令。指令生成器会随机选择场景中的物体并生成指令文本。指令文本根据指令表达模板和人工标注员创建的指代表达模板生成。指令表达模板与各指令类别相关联，总数约为 200 个；指代表达模板与各个扫描到的物体相关联，总数约为 15,000 个。

物体的指代表达分为多个详略层级生成，从最具体的（例如“白色和浅紫色分配器中的洗手液”）到最简单的（例如“它”）。选择这些表达时，要求根据表达能够从同时出现在桌子上的其他物体中唯一确定该物体。例如，如果桌子上只有一个物体，那么“把**它** 拿起来”就足以识别目标物体；但如果桌子上有许多物体，则需要更具体的表达。

为了生成通过名称指代物体的指令，我们通过爬取万维网收集了各种名称。我们收集了多种语言的人名、各类产品的品牌名称，以及昵称（例如玩具的昵称）。我们通过在指令中随机地用收集到的名称替换指代表达，生成了使用这些名称的指令。

通过组合额外表达选项（例如“你能不能”“请”），并选择是否使用冠词、是否添加修饰语，可以生成各种表达。指令生成器还可以生成负向指令（即针对不存在物体的指令、不可执行指令、不属于指令的场景描述以及随机句子）。我们希望通过使用这些负向指令进行训练，使机器人学会理解不需要执行操作的句子。通过这一流程，我们为每个场景生成了 25 条指令（总计 250,000 条）。基于使用 RefCOCOg 预训练的 MDETR 权重 [22]，我们使用生成的数据集对 I-MDETR 模型进行了微调。

附录 II 小样本目标检测

对于仿真实验，我们结合使用 Unseen Object Instance Segmentation(UOIS) [28] 和ArcFace [29]，实现小样本目标检测。UOIS 使用与 I-MDETR 相同的生成数据集进行训练。ArcFace 使用 RGB-D物体数据集 [30] 进行训练。在命名过程中，已命名物体的图像会存储在机器人记忆中；在抓取与放置指令中，UOIS 会提取桌上所有物体所在的区域，并为每个物体生成一个图像模板。每个图像模板都会与已命名物体的图像一同输入 ArcFace。这些特征会输入支持向量机 (SVM)。SVM 会预测每个模板的特征是否被归类为已命名物体所属的类别。对于真实机器人实验，我们使用在 GOT10K [31]上训练的 MixFormer [20]进行小样本目标检测。

附录 III 目标物体抓取放置方法

我们的物体抓取方法基于体积抓取网络 (VGN) [32]。我们在 MuJoCo 中实现了 Kinova Jaco2 夹爪

模型，并参考原始 URDF [23]。在仿真中生成数据集后，我们按照论文 [32] 采用监督学习训练模型。VGN 是一种为整个场景预测抓取位姿的方法，因此需要过滤掉无关的抓取位姿，才能抓取目标物体。为此，我们在仿真中使用 UOIS [28] 以及我们的小样本目标检测方法。目标物体的分割图是根据目标物体的边界框和分割图获得的，之后便可筛选抓取候选项。为了将物体放置到目标边界框中，我们通过计算边界框四个顶点位置的平均值来估计中心位置。

致谢

我们感谢 Wei Jiang 和 Noel Chen 在 Sony 实习期间实现并评估了 I-MDETR 的早期版本。我们也感谢 Hirotaka Suzuki 提供了有益的讨论。

参考文献

[1] S. Y. Min、D. S. Chaplot、P. K. Ravikumar、Y. Bisk 和 R. Salakhutdi- nov，《Film：使用模块化方法遵循语言指令》，载于国际学习表征会议，2021年。[2] D. K. Misra、J. Sung、K. Lee 和 A. S axena，《Tell me dave：将自然语言语境相关地视觉定位为操作指令》，国际机器人研究杂志，第35卷，第1–3期，第281–300页，2016年。[3] Y. Jiang、A. Gupta、Z. Zhang、G. Wang、Y. Dou、Y. Chen、L. Fei-Fei、A. Anandkumar、Y. Zhu 和 L. Fan，《Vima：通过多模态提示实现通用机器人操作》，*arXiv*预印本 *arXiv:2210 03094*，2022年。。[4] M. Shridhar、L. Manuelli 和 D. Fox，《CLIPort：机器人操作的“是什么”与“在哪里”路径》，载于机器人学习会议。PMLR，2022年，第894–906页。[5] J. Hatori、Y. Kikuchi、S. Kobayashi、K. Takahashi、Y. Tsuboi、Y. Unno、W. Ko 和 J. Tan，《通过不受限制的口语指令交互式抓取现实世界物体》，载于*2018*年*IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2018年，第3774–3781页。[6] D. Whitney、E. Rosen、J. MacGlashan、L. L. Wong 和 S. Tellex，《通过社会反馈减少物体取送交互中的错误》，载于*2017*年*IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2017年，第1006–1013页。[7] Y. Yang、X. Lou 和 C. Choi，《通过属性引导的消歧实现交互式机器人抓取》，载于*2022*年*IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2022年。[8] E. Jang、A. Irpan、M. Khansari、D. Kappler、F. Ebert、C. Lynch、S. Levine 和 C. Finn，《BC-Z：通过机器人模仿学习实现零样本任务泛化》，载于机器人学习会议。PMLR，2022年，第991–1002页。[9] S. Ishikawa 和 K. Sugiura，《Target-dependent UNITER：面向家庭服务机器人的基于Transformer的多模态语言理解模型》，*IEEE Robotics and Automation Letters*，第6卷，第4期，第8401–8408页，2021年。[10] S. Kazemzadeh、V. Ordonez、M. Matten 和 T. Berg，《ReferItGame：在自然场景照片中指称物体》，载于*2014*年自然语言处理实证方法会议论文集（*EMNLP*），2014年，第787–798页。[11] A. Kamath、M. Singh、Y. LeCun、G. Synnaeve、I. Misra 和 N. Carion，《MDETR——用于端到端多模态理解的调制检测》，载于*IEEE/CVF*国际计算机视觉会议论文集（*ICCV*），2021年10月，第1780–1790页。[12] L. H. Li*、P. Zhang*、H. Zhang*、J. Yang、C. Li、Y. Zhong、L. Wang、L. Yuan、L. Zhang、J.-N. Hwang、K.-W. Chang 和 J. Gao，《基于定位的语言-图像预训练》，载于*CVPR*，2022年。[13] T. Iio、M. Shiomi、K. Shinozawa、K. Shimohara、M. Miki 和 N. Hagita，《人机交互中的词汇协调》，国际社会机器人学杂志，第7卷，第2期，第253–263页，2015年。[14] T. Nakamura、K. Sugiura、T. Nagai、N. Iwahashi、T. Toda、H. Okada 和 T. Omori，《学习新物体以实现扩展移动操作》，Journal of Intelligent & Robotic Systems，第66卷，第1期，第187–204页，2012年。[15]《Robocup@home》，https://athome.robocup.org/，[在线；访问日期：24-February-2023]。[16] K. Zheng、X. Chen、O. C. Jenkins 和 X. E. Wang，《Vlmbench：视觉语言操作的组合式基准》，载于神经信息处理系统数据集与基准专题论文集，2022年。[17] O. Mees、L. Hermann、E. Rosete-Beas 和 W. Burgard，《Calvin：面向长时域机器人操作任务的语言条件策略学习基准》，IEEE Robotics and Automation Letters，2022年。[18] M. Shridhar、J. Thomason、D. Gordon、Y. Bisk、W. Han、R. Mottaghi、L. Zettlemoyer 和 D. Fox，《Alfred：用于理解日常任务中基于定位的指令的基准》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第10 740–10 749页。[19] J. Deng、J. Guo、N. Xue 和 S. Zafeiriou，《ArcFace：用于深度人脸识别的加性角度边际损失》，载于2019年IEEE/CVF计算机视觉与模式识别会议（CVPR），2019年，第4685–4694页。[20] Y. Cui、C. Jiang、L. Wang 和 G. Wu，《MixFormer：通过迭代混合注意力实现端到端跟踪》，载于2022年IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年，第13 598–13 608页。[21] E. Todorov、T. Erez 和 Y. Tassa，《MuJoCo：用于基于模型控制的物理引擎》，载于2012年IEEE/RSJ智能机器人与系统国际会议。IEEE，2012年，第5026–5033页。[22] J. Mao、J. Huang、A. Toshev、O. Camburu、A. L. Yuille 和 K. Murphy，《无歧义物体描述的生成与理解》，载于IEEE计算机视觉与模式识别会议论文集，2016年，第11–20页。[23]《kinovarobotics》，https://www.kinovarobotics.com/product/gen2-robots，[在线；访问日期：24-February-2023]。[24]《Intel RealSense D435i深度相机》，https://ark.intel.com/content/www/us/en/ark/products/190004/intel-realsense-depth-camera-d435i.html，[在线；访问日期：24-February-2023]。[25]《Metascan》，https://metascan.ai/，[在线；访问日期：24-February-2023]。[26] K. Mamou 和 F. Ghorbel，《三维网格近似凸分解的一种简单高效方法》，载于2009年第16届IEEE图像处理国际会议（ICIP），2009年，第3501–3504页。[27] T. Yu、D. Quillen、Z. He、R. Julian、K. Hausman、C. Finn 和 S. Levine，《Meta-World：多任务与元强化学习的基准和评估》，载于机器人学习会议（CoRL），2019 [ ] 1910 10897 。在线。可访问：https://arxiv.org/abs/。[28] C. Xie、Y. Xiang、A. Mousavian 和 D. Fox，《面向机器人环境的未见物体实例分割》，IEEE机器人学汇刊，第37卷，第5期，第1343–1359页，2021年。[29] J. Deng、J. Guo、N. Xue 和 S. Zafeiriou，《ArcFace：用于深度人脸识别的加性角度边际损失》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第4690–4699页。[30] K. Lai、L. Bo、X. Ren 和 D. Fox，《大规模分层多视角RGB-D物体数据集》，载于2011年IEEE机器人与自动化国际会议。IEEE，2011年，第1817–1824页。[31] L. Huang、X. Zhao 和 K. Huang，《GOT-10k：自然场景通用物体跟踪的高多样性大型基准》，IEEE模式分析与机器智能汇刊，第43卷，第5期，第1562–1577页，2019年。[32] M. Breyer、J. J. Chung、L. Ott、R. Siegwart 和 J. Nieto，《体积抓取网络：杂乱环境中的实时6自由度抓取检测》，载于2020年机器人学习会议论文集，机器学习研究论文集系列，J. Kober、F. Ramos 和 C. Tomlin 编，第155卷。PMLR，2021年11月16日至18日，第1602–1611页。[在线]。可访问：https://proceedings.mlr.press/v155/breyer21a.html