# Alexa Arena：以用户为中心的具身人工智能交互平台 I

Qiaozi Gao [*1] Govind Thattai [*1] Suhaila Shakiah [*1] Xiaofeng Gao [*1] Shreyas Pansare [1] Vasu Sharma [2] Gaurav Sukhatme [1] Hangjie Shi [1] Bofei Yang [2] Desheng Zheng [1] Lucy Hu [1] Karthika Arumugam [1] Shui Hu [2] Matthew Wen [1] Dinakar Guthy [1] Cadence Chung [1] Rohan Khanna [1] Osman Ipek [1] Leslie Ball [1] Kate Bland [2] Heather Rocker [1] Yadunandana Rao [1] Michael Johnston [1] Reza Ghanadan [1] Arindam Mandal [1] Dilek Hakkani Tur [1] Prem Natarajan [1]

arXiv:2303.01586v2 [cs.HC] 7 Jun 2023

## 摘要

**我们介绍了 Alexa Arena，这是一个面向用户的具身人工智能（EAI）研究模拟平台。Alexa Arena 提供多种多房间布局和可交互对象，用于创建人机交互（HRI）任务。凭借用户友好的图形和控制机制，Alexa Arena 支持开发普通人类用户即可轻松参与的游戏化机器人任务，从而为高效的人机交互数据收集和具身人工智能系统评估开辟了新途径。除该平台外，我们还介绍了一个支持对话的指令跟随基准，并提供了其基线结果。我们将 Alexa Arena[1] 公开发布，以促进通用且具备辅助能力的具身智能体构建研究。**

## 1. 引言

人工智能长期以来的目标之一，是开发能够协助人类处理日常事务的自主机器人智能体。出于多种原因，机器人实验通常在受控环境中进行，这限制了实验的多样性和规模。为缓解这一问题，近年来提出了若干具身人工智能仿真平台。这些平台支持多种虚拟场景，这些场景可以是人工设计、合成生成，或从真实场景中采集而来。具身智能体可以在这些场景中自由导航并与物体交互，以完成任务。然而，当前的 EAI 平台存在一系列局限，制约了构建可泛化的辅助型人工智能智能体的能力。

**通过游戏化促进人机交互数据收集。** 人机交互领域持续面临的一项挑战，是收集人类交互数据，包括自然语言指令以及视觉内容和动作。然而，当前可用的 EAI 平台并非为人类与智能体交互而设计。因此，数据收集过程往往成本高昂且耗时（Mandlekar 等人，2018；Srivastava 等人，2022年）。游戏历来被用于鼓励广大用户参与并积极使用各类服务（Brandle 等人，2021；Ángeles Quiroga 等人，2015年；Ram 等人，2007年）。为减轻数据收集工作，我们将游戏化引入 EAI：通过将任务或用户交互设计成游戏，并引入评分机制、成就和连续达成记录来激励用户、吸引用户参与，从而促进并更重要的是保持用户的参与。

**基于视觉观测进行推理。** 当前 EAI 平台中的任务类内变化有限：每种类型的任务都要求智能体反复使用同一物体。因此，智能体可能会采用过度设计（针对特定且简单的任务过拟合）和记忆的方法，从而收敛到琐碎且无法泛化的解决方案，或使用预定义的任务分解映射（Min 等人，2022年）。缓解这一问题的一种方法是设计这样的任务：通过与环境中的不同物体交互，并引发具有组合性且因果相连的状态变化，以多种方式完成任务。为高效完成任务，智能体需要根据视觉观测了解当前状态，并据此选择后续方案。

为此，我们提出了 Alexa Arena，一个以用户为中心的 EAI 平台。为提供更好的用户体验，该平台的设计与游戏具有共通之处，具备任务引导界面元素和引人入胜的视觉效果等特性。该平台拥有多种物体和状态转换，并在此基础上设计了各种任务（图1）。我们展示了 Arena 在对话引导的任务完成中的一个应用示例，其中具身智能体能够通过自然语言与用户交流，以完成一项室内任务。为帮助开发用于该任务的模型，我们

[*]共同贡献 [1]Amazon Alexa AI [2]此项工作是在 Amazon Alexa AI 任职期间完成的。联系：Qiaozi Gao <qz- gao@amazon.com>、Govind Thattai <thattg@amazon.com>、Xiaofeng Gao <gxiaofen@amazon.com>、Suhaila Shakiah<ssshakia@amazon.com>。

[1]https://github.com/amazon-science/alexa-arena

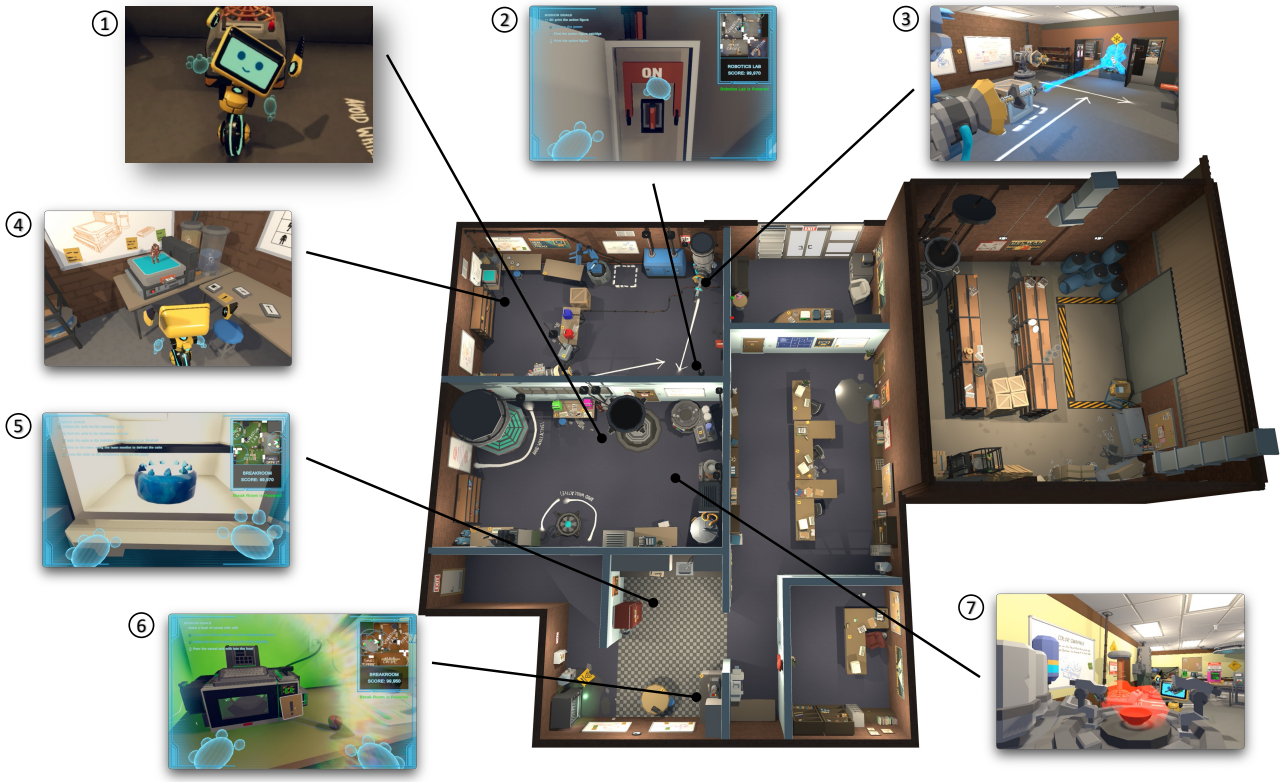



图 *1.* 作为一个 EAI 平台，Alexa Arena 拥有多种物体类别，其中包括一组用于提升参与度的奇幻物体，例如冷冻射线（3）、时光机（6）和变色器（7）。因此，智能体（1）可以利用不同工具来改变物体的状态。例如，智能体可以使用冷冻射线（3）或冰箱（5）来冷却物体。与现有平台相比，Arena 还提供了更多交互动作和状态变化。例如，智能体可以使用时光机修复损坏的物体（6），也可以使用变色器改变物体的颜色（7）。

提出了一个数据集，其中包括每项任务的专家演示，以及相应的人工标注的语言指令和对话。为实时评估对话引导的智能体，我们还发布了一个基于 Web 的用户界面（UI）作为概念验证，用户可在其中通过文本与智能体交流，并从 Arena 接收智能体的视觉观察结果。

本文的贡献如下：

1. Alexa Arena 是一个新的以用户为中心的 EAI 平台，专注于构建可泛化智能体，使其能够通过推理和程序性学习为人类提供帮助。

2. 基于 Alexa Arena 创建的对话引导任务完成基准，包含 3K 个独特任务，以及 46K 条由人类标注的指令和对话。我们还展示了基线模型在该基准上的结果。

3. 基于 Web 的用户界面，支持用户与 Arena 中的 E AI 智能体进行实时通信和交互。

## 2. 相关工作

**具身人工智能平台。** 近年来，人们提出了许多平台，让具身智能体在室内虚拟环境中执行家务活动（Puig 等人，2018；Gan 等人，2020；Xiang 等人，2020）。大多数模拟器将性能和真实性作为首要设计目标（Szot 等人，2021）。尽管一些模拟器内置了用于人工数据收集的基础设施（Gao 等人，2019；Xia 等人，2018），但用户交互通常并非其设计重点。相反，Alexa Arena 是一个以用户为中心的具身人工智能平台，专门设计了多项功能以提升用户体验：1）平台包含提升参与度的设计元素，例如奇幻物体，以及针对每项任务的评分系统；2）为了获得更好的视觉效果，Arena 中的操作以连续演绎的方式进行动画呈现，涵盖操作动作和导航动作；3）平台采用独特的用户界面，其中包含多个元素，为用户提供更好的任务引导，包括小地图、子任务提示和便笺；所有这些元素协同运作，以提升用户体验。此外，还有一些具身人工智能平台

| 平台 | 场景 | 物体类型 | 多房间 | 物体状态 | 3D Env | Real 场景 | 人类控制 | Game 就绪 |
|---|---|---|---|---|---|---|---|---|
| Malmo（Johnson 等人，2016） | 17* | 1000+ | | ✓ | ✓ | | MK | ✓ |
| Overcooked（Carroll 等人，2019年） | 6* | 5 | | ✓ | | | MK | ✓ |
| Deepmind Lab（Beattie 等人，2016） | 28* | - | | | ✓ | | MK | ✓ |
| AI2-Thor（Kolve 等人，2017） | 120 | 125+ | | ✓ | ✓ | ✓ | 鼠标和键盘、虚拟现实 | |
| iGibson（Xia 等人，2018） | 15 | 390+ | ✓ | ✓ | ✓ | ✓ | 鼠标和键盘、虚拟现实 | |
| Habitat 1.0（Savva 等人，2019年） | 1000 | - | ✓ | | ✓ | ✓ | MK | |
| Habitat 2.0（Szot 等人，2021） | 105 | 90+ | ✓ | ✓ | ✓ | ✓ | MK | |
| VirtualHome（Puig 等人，2018） | 7 | 170+ | ✓ | ✓ | ✓ | ✓ | NL | |
| TDW（Gan 等人，2020年） | 15 | 200+ | | ✓ | ✓ | ✓ | VR | |
| VRKitchen（Gao 等人，2019年） | 16 | 200+ | | ✓ | ✓ | ✓ | VR | |
| **Alexa Arena** | 10 | 335+ | ✓ | ✓ | ✓ | ✓ | MK,NL | ✓ |

*表1.* 与其他平台的比较。场景和物体：平台提供的场景数量和可交互物体类型数量。多房间：是否支持多房间场景。部分以 * 标记的平台支持自定义，并可添加新场景。状态：智能体能否通过交互改变物体状态。三维环境：平台是否支持3D虚拟环境。真实场景：虚拟环境中的场景是否逼真。人类控制：人类控制智能体的界面。鼠标和键盘：鼠标和键盘。自然语言：自然语言。虚拟现实：虚拟现实。游戏就绪功能：平台是否具有提升用户体验的游戏就绪功能，包括环境音和交互音、动作动画、流畅导航、奇幻物体、危险物、小地图、评分机制和提示。

受到电子游戏启发（Johnson 等人，2016；Beattie 等人，2016；Tian 等人，2017；Carroll 等人，2019年；Gao 等人，2020年；Yuan 等人，2022年），但其中的场景往往不够真实，且对智能体视觉感知的简化程度较高。有关 Alexa Arena 与其他具身人工智能平台的比较，请参见表 1。

**语言引导的导航与任务完成。** 现有关于学习语言引导的具身智能体的大多数研究都聚焦于导航任务（Anderson et al.，2018； Nguyen&Daumé III，2019年； Chi et al.，2020年； Roman et al.，2020年）。为应对更复杂的任务，Misra etal.，2018；Shridharetal.，2020年使智能体能够遵循自然语言指令并完成家务活动，这些活动既需要导航，也需要物体操作动作。最近，Padmakumar etal.，2022年； Gao etal.，2022年提出了新的数据集和基准，用于训练和评估能够参与对话的面向任务的具身智能体。然而，这两项研究均基于 AI2-Thor模拟器，采用离线设置，对话要么是预先从人类收集的，要么是通过模板生成的。相比之下，Arena 平台支持具身智能体以实时交互方式与人类用户交流，从而完成任务。

**使用大语言模型进行任务规划。** 近年来，使用大语言模型（LLMs）辅助机器人学习新活动或完成复杂任务的任务规划，已成为一种日益增长的趋势（Ahn 等人，2022年；Wang 等人，2023）。研究表明，大语言模型擅长提供关于物理世界和常见人类活动的高层语义知识（Huang 等人，2022a年；Singh 等人，2022年）。当与具身智能体的感知能力相结合时，这些知识可显著提升智能体在未见场景中解决复杂任务的能力（知能力后，Inoue & Ohashi，2022年；Huang 等人，2022b年）。Arena 平台为这类研究提供了良好的测试平台。凭借丰富的物体类型、属性和状态变化，以及各种环境因果事件，Arena 支持创建需要具备充分常识推理能力的机器人任务。

## 3. Arena 平台

本节将介绍 Alexa Arena 平台。我们将简要说明整体框架、模拟器的属性，以及作为平台一部分发布的工具，以便借助 Arena 启动开发。

### 3.1. 架构

Alexa Arena 框架的架构如图所示：2。平台的核心组件包括竞技场封装器和模型封装器。竞技场封装器内置竞技场协调器，它接收配置文件作为输入（详见第3.2节），并与竞技场引擎通信。竞技场引擎包含竞技场模拟器和流媒体服务器。竞技场引擎中的流媒体服务器会将机器人的第一方摄像头视图流式传输给人类用户。模型封装器包含推理模型和模型执行器，后者负责在模型与编排器之间进行接口连接。模型封装器可以修改，以开发其他模型和执行策略。此外，我们还提供了一个网页界面（即网页工具），用于

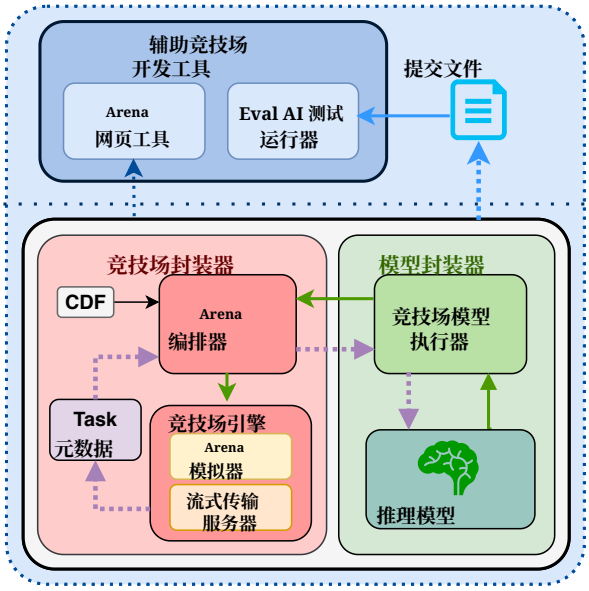



图 *2.* Alexa Arena 平台架构概览。核心模块是竞技场封装器和模型封装器。竞技场引擎包含模拟器和流媒体服务器。竞技场引擎中的元数据会发送至推理模型，推理模型生成将在模拟器中执行的动作。我们还构建了基础设施，用于为 EvalAI 挑战赛生成结果提交文件。

通过类似聊天机器人的界面与机器人模型交互，同时在网页上以流式传输的方式观看游戏进行（图3）。

### 3.2. 挑战定义格式

为定义任务配置，我们引入了一种名为挑战定义格式（CDF）的架构。CDF文件是 Arena 可读取的 JSON 格式文件，其中包含在 Arena 中配置、初始化和运行游戏任务所需的信息。这些信息包括：

**1. 游戏初始状态：**游戏场景设置、机器人初始位置、物体初始位置和初始状态（例如柜门处于打开或关闭状态）。

**2. 游戏目标条件：**需要满足的机器人和物体状态。目标条件一旦达成，即视为游戏任务已完成。

**3. 游戏相关文本数据：**用于引导用户完成游戏的文本数据，包括游戏任务描述文本、子目标描述、游戏过程中显示的系统提示等。

设计和使用 CDF 的动机，是实现大量游戏任务的程序化生成，这对于数据收集至关重要。由于人工定义的游戏任务无法扩展到为机器学习模型生成大规模训练数据，而借助易于配置的 CDF 文件，我们可以通过排列场景、机器人位置与状态、物体类型、物体位置与状态以及任务目标状态，生成数千个游戏任务。另一个优点是，由于 CDF 便于人类阅读，我们可以灵活地指定新的游戏任务，而无需编写代码。

### 3.3. 环境元数据

Arena 中的环境元数据包含机器人相机视图图像、智能体状态、物体状态、场景元数据、目标条件状态以及先前的动作执行结果。为适应不同的任务和建模设置，机器人相机视图图像支持多种模态，包括 RGB 图像、深度图像、实例分割图像和法线图像。Arena 模拟器返回的元数据响应会在每次动作执行后更新。

### 3.4. 物体属性和状态

Arena 中共有 336 种独特的物体。每种物体都具有一组属性（即可供性），用于说明是否可以进行某种类型的机器人-物体交互。例如，智能体可以切换*3D* 打印机，因为它具有可切换这一物体属性。同时，智能体无法拾取它，因为它不具有可拾取属性。总共有 14 种物体属性，包括可拾取、可打开、可破坏、容器、可切换、可通电、可弄脏、可加热、可食用、可冷却、可填充、可烹饪、装饰和可感染。每种物体属性都有对应的动作，以及执行动作后物体所处的状态。例如，打破是可破坏对应的动作，而破损则是执行该动作后的对应状态。只要满足前置条件，物体的状态就会随着动作而改变。例如，通电后（图1 （2）），*3D* 打印机便可用于制作玩具（图1 （4））。

为了进一步模拟真实世界任务和执行策略的复杂性，以及众多物体类型和属性，Arena 还提供了由常识因果事件定义的各种约束。例如，供电系统会对电器的使用施加一系列因果约束，例如，在打开灯之前需要先重置保险丝盒；微波炉需要先插入插座，才能通电启动。Arena 中的智能体也有多种方式完成任务，例如，要加热某物，智能体可以使用激光或微波炉。因此，智能体需要根据当前状态进行推理，以高效完成任务。以加热一个碗并将其送到量子实验室中的一个容器为例，此时智能体位于量子实验室。如果智能体使用激光在量子实验室中加热碗的执行路径受到物理阻碍，那么对智能体而言，更高效的做法是

前往休息室，使用微波炉更快地加热碗，然后返回量子实验室交付碗。虽然乍看之下这似乎效率较低（前往另一个房间，而不是在当前房间内用物体完成任务），但基于经验的长期推理可以赋予智能体采取高效步骤来完成任务的能力。借助因果约束、物体变化和状态转换实现的这种灵活任务完成方式，Arena 可用作情境常识推理和高级任务规划的测试平台。

### 3.5. 机器人动作空间

一般而言，Arena 支持两类动作：1）用于通过文本和视觉与用户沟通的用户交互动作；2）用于与仿真环境交互的机器人实体动作。机器人实体动作分为两种类型——导航动作和物体交互动作。为提升用户体验，所有导航和交互动作均以连续动画的方式呈现，并在动画播放期间播放环境音效。

**用户交互**。 为与用户沟通，机器人可以发起对话，对话内容会显示在用户界面上（图 3）。机器人还可以高亮显示 物体，以便实时进行视觉确认和语言指令消歧。

**导航**。 Arena 的目标是专注于组合式任务学习与推理，而非室内导航。为此，我们简化了 Arena 中的导航：机器人可以通过指定视点名称或房间名称，直接导航至房间中的某个视点；也可以通过指定物体掩码，导航至某个物体。同时，机器人也可以按需通过组合局部原子动作（例如向前移动、向后移动和旋转）逐步导航，这些动作可接受细粒度参数。为帮助机器人感知周围环境并导航至附近的物体，该平台还支持一种特殊的环视动作，使机器人能够获取全景图像。

**物体交互**。 Arena 支持 11 种物体交互动作，包括 检查、拾取、放置、打开、关闭、打破、倾倒、切换、填充、扫描 和 清洁。每种动作都对应一组处于特定状态的物体，这些物体在该状态下可执行相应动作。例如，当时光机 处于关闭 状态时，可以对其执行切换动作。交互动作会改变相关物体的状态，并在环境元数据中更新。

### 3.6. 增强用户体验

改善用户体验是 Arena 平台的一大特色，旨在启动数据收集并将人类用户纳入具身人工智能评估流程。因此，我们在 Arena 中加入了提升应用可用性、改善参与质量的元素。

1. 我们提供一种便签 物体，可在其中填写有关任务的文字提示和指南。检查 是一种特殊动作，用户可通过该动作与便签 物体交互，从中读取提示（如有需要），以完成任务目标。提示机制可作为辅助工具，向与智能体交互的用户提供数据收集相关提示，或帮助用户了解环境或任务目标（因此，这些内容不视为智能体算法或数据集的一部分）。这可以通过 CDF 进行配置。

2. 用户界面还可以在屏幕上显示目标和子目标文本，提醒用户完成任务所需完成的子目标。如果目标已完成，该功能还会显示一个勾号。这与便签不同，后者旨在引导用户完成任务。

3. 我们还在用户界面中嵌入了一个小地图，显示机器人在布局中的实时位置和朝向，以及房间名称和平面图的俯视图（图 3）。小地图还会显示闪烁信号，以吸引人类的注意并引导用户前往便笺所在的位置。

4. 该平台呈现高质量图形，支持流式传输并呈现连续动画，也就是说，打开和关闭物体，或从一个位置移动到另一个位置等动作，都会以动画形式实时、流畅地呈现给用户。此外，为了带来更引人入胜、更生动的用户体验，环境还会为某些物体状态添加动画效果，例如电源线迸出火花、热咖啡冒出蒸汽，或激光束划过空气。这些动画既适用于正在交互的物体，也适用于智能体视野内的背景物体。这使任务执行过程中的场景更加自然、易于使用且引人入胜。

5. 为了提升并保持用户参与度，我们将 Arena 上的任务游戏化。除了加入 时光机 和 变色器等有趣的奇幻物体外，我们还设置了一个显示在小地图下方的评分机制（图3）。该机制旨在实现倒计时，数值会随着时间流逝而降低。它既可激励用户取得更高分数，也可用于评估智能体。

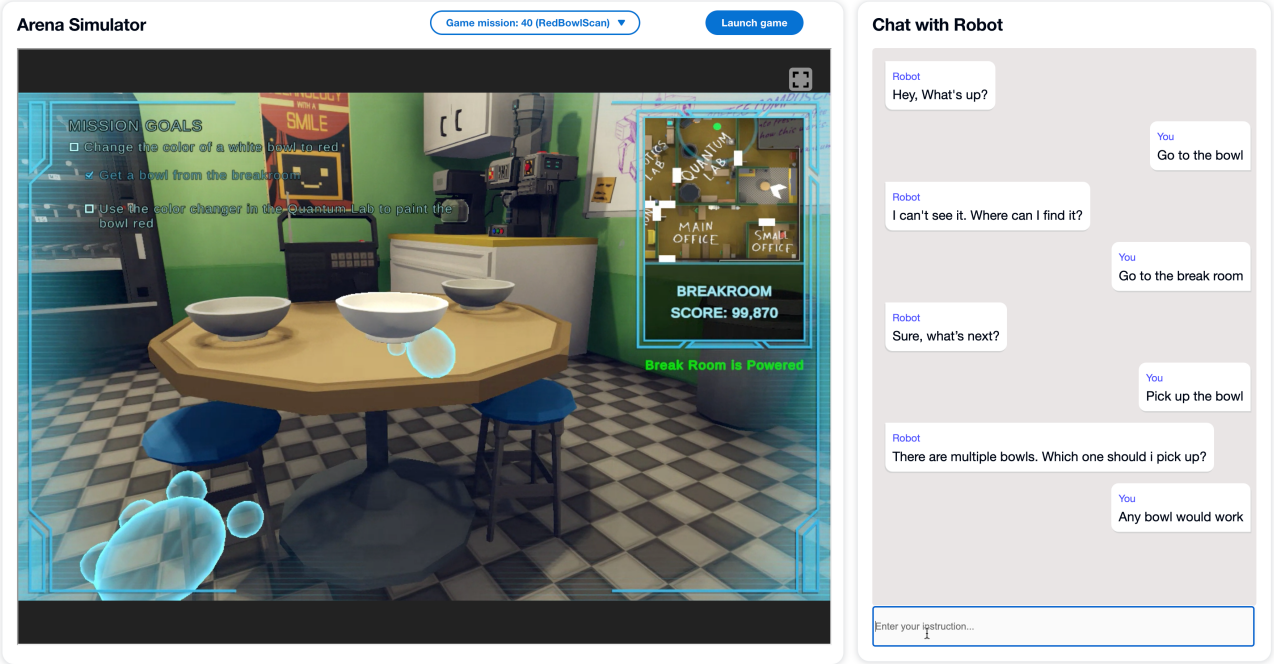



图*3.* 基于 Web 的用户界面。用户可以通过聊天框与 Arena 中的智能体交流，以完成任务。同时，Arena 的视频会实时串流，展示任务进度。小地图显示在游戏用户界面的右上角，展示房间布局以及机器人位置和朝向。界面左上角以文本形式向用户显示任务目标和子目标。

## 4. 对话引导任务完成基准

现在，我们介绍对话引导的任务完成基准。该基准旨在评估用于室内物体交互任务的对话引导智能体。为支持模型开发，我们发布了一个混合数据集，其中将真实标注的机器人动作轨迹与人类标注的语言配对。我们还在[2] EvalAI（Yadav 等人，2019年）上设立了一项挑战，并鼓励研究人员参与。

### 4.1. 任务设置

挑战中的任务要求智能体遵循用户的自然语言指令，并在虚拟环境中执行导航和物体操作动作。每个任务实例由场景的初始状态和目标完成状态，以及一系列人类指令来定义。智能体的目标是遵循人类指令，通过一系列动作与环境交互，从而达到目标状态。挑战中共有 12 种不同类型的任务，包括拾取并递送、加热并递送、冷冻并递送、修理并递送、填充并递送、着色并递送、清洁并递送、倾倒容器、打破物体、插入设备、切换设备

[2]https://eval.ai/web/challenges/challenge-page/1903/

| | # 场景 | # 任务 | # 会话 | # 指令 | # 对话 |
|---|---|---|---|---|---|
| 训练 | 6 | 2661 | 7983 | 40443 | 40105 |
| 有效 | 6 | 383 | 1149 | 6120 | 6128 |
| **总计** | 6 | 3044 | 9132 | 46563 | 46233 |

表*2.* 轨迹数据集划分。# session 表示数据会话的数量。每个数据会话对应给定任务的一条语言标注。# instr 表示每个数据划分中的指令总数。# dialog 表示问题与答案的总数。

和*scanObject*。图 4展示了几种任务类型中的数据点示例。

### 4.2. 轨迹数据集

为了支持具身智能体的训练与评估，我们收集了一个数据集，其中包含专家动作轨迹、人类语言指令，以及针对任务的问答。我们将演示和人类语言标注划分为训练集和验证集。训练集中有 2661 个任务，验证集中有 383 个任务，每个任务均由 3 名标注员标注。每条人类标注均视为一个回合，因此训练集共包含 7983 个回合，验证集共包含 1149 个回合（表2）。

| 任务 | 加热并递送 | 冷冻并递送 | 修复并递送 | 着色并递送 |
|---|---|---|---|---|
| 场景 # | 101 | 105 | 110 | 104 |
| 物体 | 碗、微波炉，desk | 马克杯、冷冻射线，桌子 | 软盘、时光机、冰箱 | 苹果，颜色变色器，书桌 |
| 指令 | 拿起碗 | 前往冷冻射线目标 | 将软盘带到时光机 | 前往颜色转换站 |
| 问题 | 碗在哪里? | 是什么<br>目标看起来像什么? | 时光机在哪里? | 在哪里车站? |
| 答案 | 在棕褐色桌子上<br>在风扇旁边 | 这是蓝色托盘<br>附着在墙上 | 时光机位于<br>在……的角落<br>台面 | 在你左侧的……<br>……的西南角<br>房间 |
| 专家演示 |  |  |  |  |

图*4.* 我们的轨迹数据集示例。每个数据会话对应于特定场景中一项待完成任务的语言标注。每项语言标注都包含一系列指令，用于引导机器人完成任务，以及用于澄清的问答。我们提供专家演示，用于开发模型。

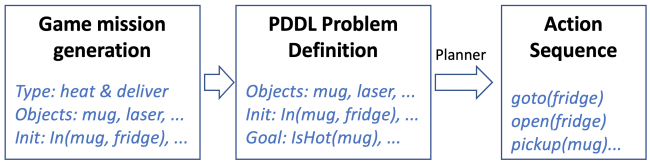



图*5.* 生成游戏任务和专家演示的流程。

#### 4.2.1. 专家演示

游戏任务与专家演示的生成流程如图 5 所示。第一步，通过从初始环境状态（例如游戏场景、机器人初始位置、与任务相关的物体及其状态）和任务目标中进行采样，以程序化方式生成游戏任务。定义游戏任务后，下一步是使用规划器生成专家演示。除了游戏定义之外，我们还为每个游戏任务生成 PDDL（规划领域定义语言）规划问题定义。符号规划器用于求解规划问题，并将输出的动作序列收集为专家演示。需要注意的是，规划器可以访问游戏元数据，而智能体在推理时无法获取这些信息。对于可以通过不同方式完成的任务（例如使用不同工具），我们为每个任务选取一种固定方式来生成专家演示。例如，对于冷冻并递送任务，在某些任务中使用冰箱冷却物体，而在另一些任务中则使用冷冻射线。

#### 4.2.2. 人类语言标注

为了收集任务的自然语言对话，我们在 Amazon Mechanical Turk（AMT）上设计了一个两阶段数据标注流程。第一阶段用于收集指令，第二阶段用于问答。在两个阶段中，标注员都会观看包含该任务专家演示的视频，并以自由文本数据的形式提供标注，或回答选择题。

在第一阶段，标注者需要编写指令，告诉“智能机器人”如何完成任务。在此过程中，标注者首先观看真实机器人动作的视频，然后为每个高亮显示的视频片段编写指令。图 9a 展示了用于收集语言指令的用户界面。

收集完所有指令后，我们开始标注第二阶段。在这一阶段，标注员需要提出问题，以便更好地完成任务，就像他们在控制机器人按照指令行动一样。随后，他们还需要回答自己提出的问题。与（Gao 等人，2022年）类似，标注第二阶段的问题选项由预定义模板生成。给定一条指令，我们会提取其中的名词并填入模板，以构造问题。提取出的名词对应于智能体为完成任务而需要进一步了解的查询对象。有关人类语言标注过程的更多详情，请参阅附录A.1.1。




图 *6*。视觉数据集概览。该数据集由 RGB 图像和 PNG 格式的真实分割图像组成，每张图像还配有一个包含元数据信息的 JSON 文件。

### 4.3. 视觉数据集

视觉感知对于具身人工智能智能体在环境中导航和交互至关重要。具备视觉智能的机器人利用视觉线索设定导航目标，并根据感知到的物体状态变化构建视觉表征。为了推动具身人工智能相关领域内外的视觉人工智能研究，我们发布了一个基于 Arena 的视觉数据集。该数据集包含 60 万张训练图像和 6 万张验证图像，涵盖 336 种独特物体，这些物体分属 160 多个语义分组。该数据集可用于 Arena 上的大规模计算机视觉研究，例如物体检测、物体状态分类等。数据集的整体结构和内容如图6所示。视觉数据集通过以编程方式配置 CDF，并在 Arena 环境中初始化包含所有可生成物体的场景来收集。随后，我们导航至每个物体的位置，从不同视角和距离拍摄图像。视觉数据生成过程的详细说明见附录A.1.4。

## 5. 基线模型

我们提供了两个基线模型，分别在轨迹数据集和视觉数据集上训练，并在验证轨迹数据集上进行评估。模型的输入是智能体的第一人称视角和自然语言指令。模型的输出是一系列动作序列（以及必要时附带的可交互对象掩码），供智能体在环境中执行，以完成当前任务。

### 5.1. 神经符号方法

Episodic Transformer（ET）是一种用于视觉语言导航的多模态 Transformer 模型（Pashevich 等人，2021）。ET 利用整个回合中的历史视觉和语言信息，捕捉动作之间的长期依赖关系。它最初是为 Alfred 挑战赛设计的（Shridhar 等人，2020年）。我们将 Episodic Transformer 模型应用于我们的挑战赛，并使用在 Arena 数据集上训练的视觉模型进行物体检测。其架构如图7a所示。

#### 5.1.1. 视觉模型

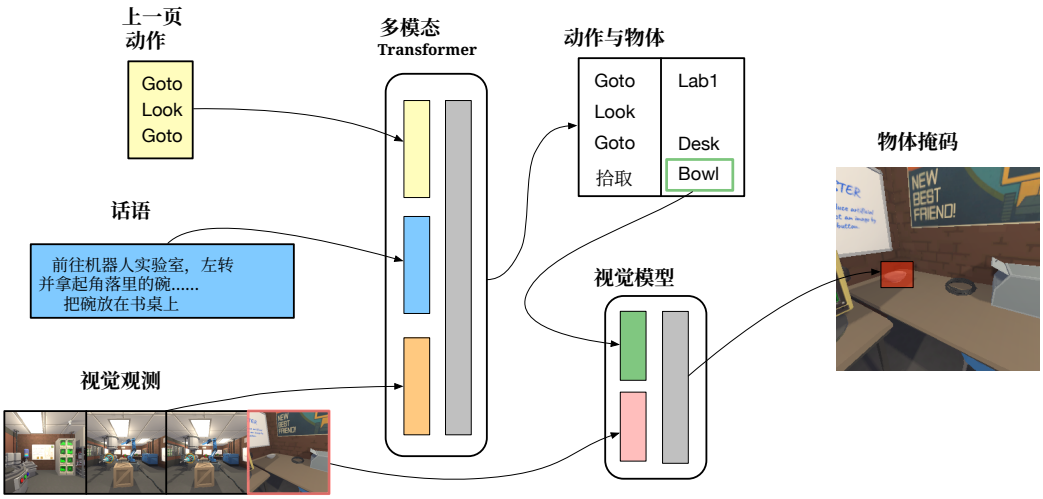
我们使用视觉数据集训练 Mask-RCNN 图像分割模型（He 等人，2017）。我们对视觉数据集中的物体类别进行处理，将其归并为 86 个语义类别，其中包括背景类别。该模型以 RGB 图像作为输入，并预测图像中所有物体实例（涵盖 86 个物体类别）的掩码，以及相应的类别标签和置信度分数。随后，将掩码和物体类别预测结果与多模态 Transformer 预测的动作和物体结合，得到可交互对象的单个掩码预测。

视觉模型使用 SGD优化器训练 22 轮，学习率为 0.00125，权重衰减为 0.001；在训练 15 轮后，学习率按 0.1 的系数衰减。我们使用全局批次大小 16，并在 2 块 Tesla V100 GPU 上训练模型。

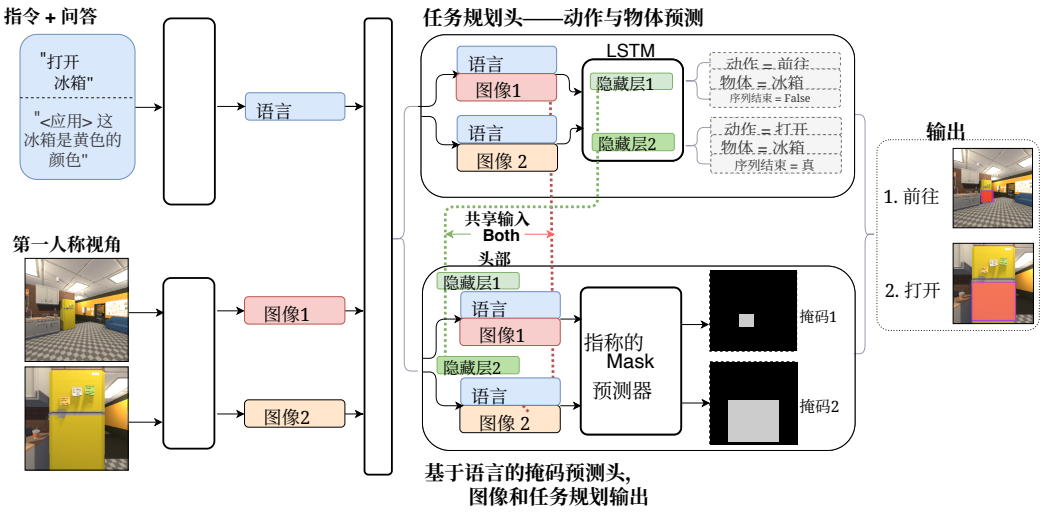
#### 5.1.2. 实验设置

多模态 Transformer模型使用先前的动作、先前的视觉观测以及该回合的完整话语，在预测出作为动作的 停止 标记之前，预测下一个动作类型和物体类别。对于视觉观测，我们首先使用在视觉数据集上训练的视觉模型的骨干网络提取特征，然后将其输入 Transformer。训练期间，每次执行环顾动作后，模型都会利用来自不同方向的观测来预测旋转次数；旋转次数取决于专家演示中后续动作所使用的四幅全景图像中的哪一幅。推理期间，视觉模型会检测当前图像中的所有物体，并使用预测出的物体类别生成相应物体的掩码。请注意，如果视觉模型检测到同一类别的多个物体实例，我们会选择似然度最高的实例。

对于语言话语，我们尝试仅使用人类指令作为输入，也尝试将指令与问题和答案一同作为输入。我们使用两个词元来高效编码预定义问题，包括 (i) 表示问题类型的词元（即，位置类问题询问物体位置，外观类问题询问物体外观，方向类问题询问方向，指代类问题询问指代），以及 (ii) 在位置、外观和指称类问题中表示目标物体的词元。例如，询问微波炉外观的问题将编码为 外观 微波炉。由于方向问题询问的是方向，与物体无关，因此不需要第二个词元，这类问题中会省略该词元。对于不属于预定义




(a) 神经符号方法




(b) 视觉语言模型

图 7。模型架构。a) 神经符号方法使用彼此独立的模块来预测动作和掩码，并使用启发式规则进行视觉定位。b) 视觉语言模型以端到端的方式进行动作预测和神经视觉定位。

类型的问题，我们会纳入完整的问题和答案。

我们使用批量大小为 2 的设置训练 Transformer 模型 20 个训练轮数。我们使用 Adam 优化器，前 10 个训练轮数的学习率为 $1e^{-4}$，其余训练轮数的学习率为 $1e^{-5}$。

## 5.2. 端到端视觉语言模型

我们尝试使用端到端视觉语言（VL）模型，根据自然语言指令和智能体的第一人称视角预测动作和可交互对象掩码。该模型利用自然语言指令中包含的消歧信息进行视觉定位，即为可交互对象实例预测单个掩码，而不是采用基于规则的定位。

### 5.2.1. 实验设置

模型的架构如图所示 7b。高层级自然语言指令使用 Transformer 模型进行编码。为了展示对话诱导的问题与答案（问答）的作用，我们分别尝试仅编码语言指令（不含问答），以及编码语言指令、问题类型和相关问题的答案，以纳入上下文和消歧信息（含问答）。我们使用与 ET 模型相同的词元来编码问题类型。自我中心 RGB图像使用 ResNet（He 等人，2015年）或 ViT（Dosovitskiy 等人，2020年）骨干网络进行编码。我们的实验采用 ResNet 骨干网络。图像编码器和语言编码器使用来自 CLIP 的预训练权重进行初始化

（Radford 等人，2021）。随后，视觉特征和语言特征被输入一个联合视觉语言编码器，该编码器输出组合后的视觉语言嵌入。接着，嵌入被送入两个头部：

1. 任务规划头将编码后的 VL 嵌入输入由两层组成的自回归长短期记忆网络模块。该组件对先前预测输出的历史上下文进行编码，然后通过两个线性分类器，预测与输入语言和图像特征相关联的*(*动作、物体*)* 对。

2. 掩码预测头利用当前 RGB 第一人称视角、编码后的语言，以及任务规划头输出的动作—物体预测隐藏状态，预测一个指向待交互物体的掩码。我们采用 (Wang 等人，2022年) 中描述的指代图像分割架构，为自然语言指令所指向的可交互物体实例预测单个掩码。

针对一条自然语言指令和一系列第一人称视角，重复上述过程，直至任务规划头发出序列结束（EOS）标志（该标志由每个时间步对应的独立二元分类器建模）。

训练时，我们为骨干网络和掩码预测头使用 $1e^{-6}$ 的学习率，为动作预测头使用 $1e^{-4}$ 的学习率。我们使用 Adam 优化器，权重衰减为 $1e^{-5}$，并在训练 35 轮后将学习率衰减至 0.1。模型在 4 个 Tesla V100 GPU 上训练 65 轮，全局批次大小为 64 （每个 GPU 16 ）。

## 6. 结果与讨论

我们在轨迹数据集上展示结果。模型在训练集划分上进行训练，并在验证集划分上进行评估。

### 6.1. 评估指标

#### 6.1.1. 智能体评估

我们期望优秀的具身智能体能够高效地完成任务。为实现这一目标，智能体应理解人类指令并生成相应的动作序列。因此，我们使用以下指标评估智能体：

• **任务成功率（MSR）**。对于每项任务，都有一个任务成功变量 $m$，用于指示该任务的目标条件是否已满足。 $m = 1$ 如果所有目标条件均已满足。任务完成率通过对所有任务的 $m$ 取平均值计算得出。

• **机器人平均动作数（NRA）**。为了衡量任务完成效率，我们还记录智能体完成每项任务所采取的动作数量。机器人平均动作数通过对所有任务中的动作数量取平均值计算得出。

#### 6.1.2. 视觉模型评估

神经符号模型由一个独立的视觉组件构成，该组件对智能体的第一人称视角和/或周围图像执行实例级分割。我们使用标准的 COCO 评估指标——平均精度均值（mAP）——评估实例分割模型。mAP 指标通过计算交并比（IoU）阈值从 0.5 到 0.95、步长为 0.05 时的精度均值获得。预测实例类别的分数不纳入考量，而是对每张图像的最大检测数量设置上限（设为 100）。为了给端到端模型选择工作点并针对召回率进行调优，除了标准的 COCO 平均精度均值指标外，我们还提供一种基于分数和 IoU 阈值的精度指标，称为 *t-mAP*。该指标通过将输出实例类别分数阈值设为 [0.05、0.1、0.3、0.5、0.7], , 并将 IoU 阈值设为 [0.1、0.3、0.4、0.5、0.75、0.8] , 对全部 30 种（分数、IoU）组合进行阈值筛选后取平均值。我们同时报告所有物体的两项指标，以及按小型、中型和大型物体分类的指标。结果见表 3a。

### 6.2. 结果与分析

基线模型的总体任务完成结果显示在表3b 和表3c中。两个模型的评估均将每项任务允许的最大步数上限设为 50 步，并将每项任务允许的失败步数上限设为 10 步；超过该上限后，如果任务目标仍未完成，智能体将停止执行，使目标保持未完成状态。下面，我们将分析两个模型的结果，并深入探讨模型性能。

#### 6.2.1. 神经符号模型

总体而言，加入问答后，神经符号模型的性能略有提升。观察各类任务的 MSR（表3c），可以发现，模型在时间跨度较短的任务（例如 scanObject、toggleDevice）上表现良好。作为一项消融研究，我们还在验证集上评估了多模态 Transformer，向其提供每个时间步的真实视觉观测。结果表明，对于 71.3% 的任务，模型在没有问答的情况下能够正确预测所有动作和物体（包括旋转动作）；加入问答后，这一比例为 76.9%。这些定量结果表明，神经符号模型的瓶颈

| 类别 | 面积（$px^2$） | COCO 平均精度均值 t-mAP | |
|---|---|---|---|
| 小 | 0 - 1296 | 37.63 | 91.49 |
| 中 | 1296 - 9216 | 60.41 | 92.15 |
| 大 | 9216 - 90000 | 64.72 | 84.91 |
| 总体 | 0 - 90000 | 46.03 | 89.5 |

表3a. 神经符号方法中所用视觉模型的图像分割结果。我们展示了小型、中型和大型物体的各项指标。

| 方法 | MSR (%) 机器人动作平均次数 | |
|---|---|---|
| 神经符号 | 18.19 | 11.82 |
| 视觉语言模型 | 22.80 | 12.73 |
| 神经符号（含问答） | 19.32 | 11.73 |
| 视觉语言模型（含问答） | **34.20** | **8.82** |

表3b. 两种基线方法在验证数据集上的智能体评估实验结果。MSR 表示任务成功率。NRA 表示机器人的平均动作次数。

| 任务 Type | NS (%) | 神经符号（%） | 视觉语言（%） | 视觉语言（%） |
|---|---|---|---|---|
| | 不含问答 | 含问答 | 不含问答 | 含问答 |
| 打破物体 | 0.00 | 0.00 | 21.11 | 41.11 |
| 清洁并递送 | 12.64 | 13.79 | 13.79 | 19.10 |
| 着色并递送 | 0.00 | 0.00 | 0.00 | 0.00 |
| 填充并递送 | 14.58 | 18.75 | 10.41 | 22.91 |
| 冷冻并递送 | 33.33 | 25.00 | 0.00 | 8.33 |
| 加热并递送 | 5.13 | 5.13 | 10.25 | 28.20 |
| 插入设备 | 14.12 | 14.69 | 14.68 | 20.90 |
| 拾取并递送 | 9.47 | 12.63 | 15.43 | 27.36 |
| 倾倒容器 | 14.53 | 16.24 | 16.23 | 30.76 |
| 修理并交付 | 11.11 | 12.96 | 9.25 | 29.62 |
| scanObject | 41.44 | 41.44 | 37.83 | 56.75 |
| 切换设备 | 57.14 | 56.19 | 81.90 | 81.90 |

表3c. 基于任务类型的神经符号（NS）和视觉语言（VL）模型的 MSR。结果分别为使用问答和不使用问答的情况。

表*3.* 基线模型评估结果

模型并非负责预测动作和物体。事实上，由于一个房间中常常有多个属于同一类别的物体，即便 Transformer 模型正确预测了物体类别，视觉模型也正确预测了图像中该类别的所有物体掩码，但由于我们使用启发式方法进行视觉定位，模型在为正确的物体实例选择掩码时仍可能出错（图 8）。这也可以解释为什么问答带来的改进有限，因为一旦预测出动作和物体，语言指令就不再参与定位。此外，视觉模型难以检测某些小型物体，例如胡萝卜、咖啡杯、锤子和螺丝刀，这也会造成一些错误。有些物体看起来相似，却属于不同类别，例如时光机和微波炉、桌子和书架，这使得视觉模型难以区分它们。

6.2.2. 视觉语言模型

与未使用问答进行训练和评估相比，视觉-语言模型在使用问答进行训练和评估时，MSR 绝对提升了 11.40%。使用问答后，智能体也能更高效地完成任务，任务完成的平均步数减少了 3.91 步。这表明，提出正确的问题能够提升任务执行效率和准确性。对于问答能够提供有价值信息的任务类型，智能体的 MSR 显著提升。对于 切换 这类只有少数物体可以执行的任务，指令通常已经以自然语言进行了足够清晰的描述（“打开红色计算机”，“按下蓝色按钮”），因此添加问答不会带来任何提升。对于所有其他涉及众多物体的任务类型，问答能够提供指令中通常不会自然包含的关键额外信息，从而显著提升性能。分析表明，颜色 &



图*8*。图像中有两张桌子，左侧是一张蓝色的桌子，右侧是一张绿色的桌子。如果没有得到充分指引，智能体将面临具有挑战性的消歧任务。

递送任务容易失败，因为难以明确递送的容器。

尽管视觉-语言模型通过同时生成动作和掩码的端到端方法展现出更好的性能，但通过分析机器人在 25 个任务中的轨迹，我们总结了端到端架构的关键见解、错误来源和失效点：

1. 有些任务失败，是因为在存在多个实例候选项时，无法将自然语言指令在视觉上定位到单个可供交互的物体实例。例如，在图8中，有两张书桌——左侧是蓝色的，右侧是绿色的。如果自然语言指令描述得不够清楚，模型就无法在视觉上定位感兴趣的书桌。分析后我们发现，在 *<x>&deliver* 任务类型中，智能体通常能够完成 *<x>* 目标，却会因为语言存在歧义（*deliver* 目标（例如“书桌就在你正前方，上面放着显示器”）而失败；模型的视觉感知也可能无法区分同一物体的多个实例。

2. 有时，智能体会因错误预测而偏离路径（例如，进入错误的房间），随后要求它在房间里寻找物体的指令，会使它陷入无限旋转、试图寻找该物体的循环；也就是说，智能体试图响应下一条指令，而这条指令假设前一个状态已成功达成。模型可以通过更好的错误纠正机制来识别这类循环状态，并跳出循环探索轨迹中的其他选项，或者执行策略回溯到先前的指令，从而有望解决这一问题。由于当前的基线模型使用数据集中现成的答案，因此也可以改进模型，使其能够根据实时的当前视觉观测提出问题。

3. 由于这是一个端到端模型，语言和视觉特征会被联合编码，并通过端到端训练，从输入语言和图像映射到输出掩码和动作。有些错误是因为模型更关注眼前的视觉信息，而不是语言输入。例如，如果机器人站在柜门或放有咖啡壶的架子旁边，即使语言指令是“前往激光炮处”，它仍会分别尝试打开 和关闭 柜门，或拾取咖啡壶；这些行为都是模型从数据中学到的。

4. 模型在所有评估任务中还存在一些其他误差来源，包括语言歧义、无法对长程任务指令进行建模，例如“右转，前往休息室，加热碗并将其放在桌子上”——该指令至少包含 9 个步骤；无法检测某些极小物体或距离过近的物体的掩码；以及无法区分外观极其相似的物体（例如激光炮和冷冻射线）。这些是神经符号模型和视觉-语言模型共有的问题。

5. 我们还进行了消融研究：仅训练掩码预测头，并使用真实动作预测结果。在验证数据集上评估该模型，得到的 MSR 为 71.89%。这模拟了这样一种情形：预测出完美的动作，使机器人处于能够为可交互对象预测掩码的位置，从而仅评估物体检测以及对自然语言指令的视觉落地能力。该结果表明，模型难以端到端地找到物体位置，因此需要纳入更好的路径规划模块，使其能够从多模态输入中学习。

## 7. 结论与未来工作

本文介绍了 Alexa Arena，这是一个以用户为中心的具身人工智能平台。Alexa Arena 具有用户友好的图形、动画和控制机制，可促进大规模、以用户为中心的具身人工智能研究。Alexa Arena 以游戏形式面向普通用户呈现，其中的任务涵盖多种经过精心设计的物体类别和交互动作，为数据收集、具身人工智能开发和评估开辟了新的可能性。我们展示了该平台在对话引导的任务完成方面的一个用例，并发布了一个混合数据集，其中包含任务规划、视觉定位和自然语言理解方面的挑战性问题。我们针对两种模型架构，在对话增强设置下提供了基线结果。未来，我们旨在引入一种基于模型的提问机制，在任务执行过程中由预言机提供相关答案，从而形成更准确、更具动态性的模型，利用实时上下文信息进行定位。另一个

方向是使人类和机器人能够控制不同的具身智能体，并为人机交互创造更多场景。我们还计划在未来更新中发布更多任务和场景，以将其他令人兴奋的元素纳入Arena，例如重力翻转器、危险物以及智能体执行路径上的其他阻挡物，从而为多模态推理和组合式任务规划的创新开辟途径。

## 参考文献


Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., 等。按我能做的来做，而不是照我说的做：将语言落地于机器人可供性。*arXiv* 预印本 *arXiv:2204.01691*，2022年。

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. 视觉与语言导航：在真实环境中解读基于视觉的导航指令。载于*IEEE*计算机视觉与模式识别会议论文集，第3674–3683页，2018。

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright,M., Küttler, H., Lefrancq, A.,Green,S., Valdés, V., Sadik, A., 等。DeepMind Lab。*arXiv* 预印本 *arXiv:1612.03801*，2016。

Brandle, F., Allen, K., R, Tenenbaum, J., and Eric, S. 运用游戏理解智力。认知科学学会年会论文集，*43*，2021。

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. 关于学习人类对人类与人工智能协同的效用。神经信息处理系统进展，32，2019年。

Chi, T.-C., Shen, M., Eric, M., Kim, S., and Hakkani-tur, D. 直接提问：用于视觉与语言导航的交互式学习框架。载于*AAAI*人工智能会议论文集，第34卷，第2459–2466页，2020年。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等。图像胜过千言：用于大规模图像识别的Transformer。*arXiv*预印本 *arXiv:2010.11929*，2020年。

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., 等。Threedworld：用于交互式多模态物理仿真的平台。*arXiv*预印本 *arXiv:2007.04954*，2020年。

Gao, X., Gong, R., Shu, T., Xie, X., Wang, S. 和 Zhu, S.- C. VRKitchen：用于面向任务的学习的交互式3D虚拟环境。*arXiv*预印本 *arXiv:1903.05757*，2019年。Gao, X., Gong, R., Zhao, Y., Wang, S., Shu, T. 和 Zhu, S.-C. 复杂人机协作任务中的解释生成联合心智建模。载于*2020*年第*29*届*IEEE*机器人与人类交互通信国际会议 *(RO-MAN)*，第1119–1126页。IEEE，2020年。

Gao, X., Gao, Q., Gong, R., Lin, K., Thattai, G. 和 Sukhatme, G. S. Dialfred：用于具身指令遵循的支持对话的智能体。*arXiv* 预印本 *arXiv:2202.13330*，2022年。

He, K., Zhang, X., Ren, S. 和 Sun, J. 用于图像识别的深度残差学习，2015年。网址 https://arxiv.org/abs/1512.03385。

He, K., Gkioxari, G., Dollár, P. 和 Girshick, R. B. Mask R-CNN。*CoRR*，abs/1703.06870，2017。

Huang, W., Abbeel, P., Pathak, D. 和 Mordatch, I. 将语言模型用作零样本规划器：为具身智能体提取可执行知识。发表于国际机器学习会议，第 9118–9147 页。PMLR，2022a年。

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y. 等。《Inner Monologue：通过语言模型规划实现具身推理》。*arXiv*预印本 *arXiv:2207.05608*，2022b年。

Inoue, Y. 和 Ohashi, H.《Prompter：利用大语言模型提示，实现数据高效的具身指令遵循》。*arXiv*预印本 *arXiv:2211.03267*，2022年。

Johnson, M., Hofmann, K., Hutton, T. 和 Bignell, D.《用于人工智能实验的 Malmo 平台》。发表于*IJCAI*，第 4246–4247 页。Citeseer，2016。

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A. 和 Farhadi, A.《AI2-THOR：用于视觉人工智能的交互式三维环境》。*arXiv*预印本 *arXiv:1712.05474*，2017。

Mandlekar, A., Zhu, Y., Garg, A., Booher, J., Spero, M., Tung, A., Gao, J., Emmons, J., Gupta, A., Orbay, E. 等。RoboTurk：一个通过模仿进行机器人技能学习的众包平台。载于 机器人学习会议，第 879–893 页。PMLR，2018。

Min, S. Y., Chaplot, D. S., Ravikumar, P. K., Bisk, Y., and Salakhutdinov, R. FILM：使用模块化方法遵循语言指令。载于国际学习表征会议，2022年。

Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., and Artzi, Y. 利用视觉目标预测将指令映射为三维环境中的动作。*arXiv*预印本 *arXiv:1809.00786*，2018。

Nguyen, K. 和 Daumé III, H. Help, anna!：通过回顾式好奇心激励模仿学习，借助自然多模态辅助进行视觉导航。载于自然语言处理经验方法会议（*EMNLP*）论文集，2019年11月。

Padmakumar, A.、Thomason, J.、Shrivastava, A.、Lange, P.、Narayan-Chen, A.、Gella, S.、Piramithu, R.、Tur, G. 和 Hakkani-Tur, D. TEACh：任务驱动的对话式具身智能体。载于人工智能会议（*AAAI*），2022年。

Pashevich, A.、Schmid, C. 和 Sun, C. 用于视觉与语言导航的 Episodictransformer。*2021 IEEE/CVF*国际计算机视觉会议（*ICCV*），第15922–15932页，2021。

Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., 和 Torralba, A. Virtualhome：通过程序模拟家务活动。载于*IEEE*计算机视觉与模式识别会议论文集，第8494–8502页，2018。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., 和 Sutskever, I. 从自然语言监督中学习可迁移视觉模型，2021。

Ram, A., Ontañón, S., 和 Mehta, M. 用于自适应电子游戏的人工智能。载于佛罗里达人工智能研究学会，2007年。

Roman, H. R., Bisk, Y., Thomason, J., Celikyilmaz, A., 和 Gao, J. Rmm：用于对话导航的递归心理模型。载于*2020*年自然语言处理经验方法会议论文集，2020年。

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., 等。Habitat：具身人工智能研究平台。载于*IEEE/CVF*国际计算机视觉会议论文集，第9339–9347页，2019年。

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., 和 Fox, D. Alfred：用于解读日常任务中基于环境的指令的基准。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第10740–10749页，2020年。

Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J. 和 Garg, A. Prog-prompt：利用大语言模型生成情境化机器人任务规划。*arXiv* 预印本 *arXiv:2209.11302*，2022年。

Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K. E., Lian, Z., Gokmen, C., Buch, S., Liu, K. 等。Behavior：虚拟、交互式和生态环境中的日常家务活动基准。在机器人学习会议，第 477–490 页。PMLR，2022年。

Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D. S., Maksymets, O., 等。Habitat 2.0：训练家居助手重新布置其栖息环境。神经信息处理系统进展，34：251–266，2021。

Tian, Y., Gong, Q., Shang, W., Wu, Y., and Zitnick, C. L. Elf：一个面向即时战略游戏的全面、轻量且灵活的研究平台。神经信息处理系统进展,30, 2017.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and Liu, T. Cris：由 CLIP 驱动的指代表达图像分割。载于*IEEE/CVF*计算机视觉与模式识别会议论文集, 2022.

Wang, Z., Cai, S., Liu, A., Ma, X., and Liang, Y. 描述、解释、规划与选择：大语言模型支持的交互式规划使开放世界多任务智能体成为可能。*arXiv*预印本 *arXiv:2302.01560*, 2023.

Xia, F., Zamir, A. R., He, Z., Sax, A., Malik, J., and Savarese, S. Gibson Env：面向具身智能体的现实世界感知。载于*IEEE*计算机视觉与模式识别会议论文集, pp. 9068–9079, 2018.

Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Ji ang, H., Yuan, Y., Wang, H., Yi, L., Chang,A. X., Guibas, L. J. 和 Su, H. SAPIEN：一种基于部件的交互式仿真环境。2020年6月。

Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., Singh, S. B., Lee, S. 和 Batra, D. Evalai：面向更完善的人工智能智能体评估系统。*arXiv* 预印本 *arXiv:1902.03570*，2019年。

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., Lu, H., Zhu, Y. 和 Zhu, S.-C. 原位双向人机价值对齐。《科学机器人》，7(68):eabm4183，2022年。

Ángeles Quiroga, M., Escorial, S., Román, F. J., Morillo, D., Jarabo, A., Privado, J., Hernández, M., Gallego, B., 和 Colom, R. 我们能通过商业电子游戏可靠地测量智力一般因子（g）吗？答案是肯定的！《智力》，53:1–7，2015年。ISSN 0160-2896。doi: https://doi.org/10.1016/j.intell.2015.08.004。

# A. 附录

## A.1. 数据收集

### A.1.1. 人类语言标注

为了收集任务中的自然语言对话，我们设计了一个分两个阶段的数据标注流程，并在 Amazon Mechanical Turk（AMT）上进行。第一阶段用于收集指令，第二阶段用于问答。在两个阶段中，标注员都会观看一段包含该任务专家演示的视频，并通过文本或选择题提供标注。

在第一阶段，标注者需要编写指令，告诉“智能机器人”如何完成任务。在此过程中，标注员会先观看真实机器人动作的视频，然后针对每个突出显示的视频片段编写指令。图 9a展示了用于收集语言指令的用户界面。

收集完所有指令后，我们开始标注第二阶段，要求标注者提出问题，以帮助自己完成任务，就像他们正在控制机器人按照指令行动一样。他们还需要回答自己提出的问题。图 9b 展示了用于收集问答的界面。标注者需要遵循以下步骤：

1. 标注员观看视频：播放一段 10 秒的视频片段，直到子任务开始，并在屏幕上显示相应的指令。该视频片段有助于人类标注员理解初始环境状态。

2. 标注员选择问题：标注员选择一个认为可能有助于完成任务的相关问题。这些问题由若干预定义模板生成。标注员也可以选择输入自己的问题。

3. 标注员回答问题：同一位标注员观看专家智能体按照语言指令执行指定子任务的视频片段，然后以文本形式回答问题。

4. 标注员指出是否有必要提问：有时，指令和视觉上下文已经足够清晰，因此没有必要提问。

为确保标注质量，在实际操作中，我们在第一阶段的数据收集中采集了 3000 个数据会话。随后，我们请其他标注者找出并尝试纠正标注错误，例如错误的语言指令、语法错误、指令与机器人动作不匹配等。对于大多数采集到的标注都没有意义且不值得修正的数据会话，我们要求其他标注者直接忽略。最终，20.3% 的数据会话被忽略，35.7% 的指令得到修正。根据数据验证结果，我们筛选出一批生成了高质量数据的 AMT 工作人员（约 300 人），并在后续的数据收集过程中继续使用他们。

### A.1.2. 问题生成

与（Gao 等人，2022年）类似，标注第二阶段的问题选项使用预定义模板生成。给定一条指令，我们提取名词并将其插入模板中以构造问题。提取的名词对应于智能体为完成任务而需要进一步了解的查询对象。

具体而言，我们考虑三种问题类型，分别涉及查询物体 $o$ 的位置和外观，以及智能体与目标位置之间的相对方向。每种问题的模板定义如下：

1. 位置： $o$ 在哪里?

2. 外观： $o$ 看起来是什么样子?

3. 方向：我应该转向哪个方向?

4. 指代：你指的是哪个 $o$？

图 10 和图 11 展示了轨迹数据集的统计信息。

### A.1.3. 预言机答案

人类语言具有很大的多样性和复杂性，因此智能体理解起来可能颇具挑战。作为替代方案，我们可以先向智能体提供模板化语言，作为其进行语言理解的起点。为此，除了请人工标注者回答问题外，我们还构建了一个预言机，通过从模拟器元数据中提取信息来回答这些问题：(i) 为回答位置问题，我们计算物体相对于智能体的方向，以及房间中距离物体最近的视点。为了帮助检测小型物体，如果物体由容器持有，我们还提供该容器以及同一容器中持有的其他物体的信息，作为地标；(ii) 为回答外观问题，我们直接从模拟器元数据中提取物体的形状、颜色和材料信息；(iii) 对于方向问题，我们比较智能体在子任务结束时与初始位置的位置，以提供智能体移动方向的答案。获得必要信息后，我们使用语言模板生成答案。模板示例包括：

| 任务类型 | 训练 | 有效 |
|---|---|---|
| 打破物体 | 196 | 30 |
| 清洁并递送 | 161 | 29 |
| 着色并递送 | 29 | 4 |
| 填充并递送 | 108 | 16 |
| 冷冻并递送 | 76 | 8 |
| 加热并递送 | 42 | 13 |
| 插入设备 | 381 | 59 |
| 拾取并递送 | 669 | 95 |
| pour | 265 | 39 |
| 修理并交付 | 131 | 18 |
| scanObject | 324 | 37 |
| 切换设备 | 279 | 35 |

表 4a. 轨迹数据集训练集和验证集划分中任务类型的分布。

1. 位置： *o* 位于你的 [方向] 的 [容器] 中，在 [地标] 旁边，位于 [房间]内。它距离 [视点]最近。 2. 外观： *o* 是 [形状] 的，颜色为 [颜色]。它由 [材料]制成。 3. 方向：你应该向 [方向] 转身/你无需移动。

#### A.1.4. 视觉数据生成

图像数据收集策略简单直接——目标是从不同距离、视图和视角，收集涵盖 Arena 中所有可能生成的物体的数据。具体步骤如下：

1. 使用包含所需物体的 CDF文件初始化游戏任务。 2. 初始化后，我们发出命令，使用基本导航动作前往场景中的所有不同物体，并捕捉第一人称视角。3. 随后，我们根据捕捉时该物体的可供性发出交互指令来捕捉图像，这些可供性取自游戏元数据响应。4. 我们还使用基本三角学在物体周围进行随机移动，以获取额外视角。

针对不同的照明条件，我们在灯光开启和关闭的情况下，使用各种任务和物体重复上述步骤。所有命令、物体状态及其在场景中的位置，以及大量其他有用信息，都会记录在元数据文件中。我们为每张图像提供 RGB彩色图像、分割图和相关元数据。数据集的统计信息如图12b所示。

**数据集验证**。为验证用于建模的真实值分割和边界框标注的质量，我们收集了 1245张图像，其中包含涵盖所有物体类别的实例，并为其标注边界框及对应的类别名称。由于 Arena 中有奇幻物体（如重力翻转器），我们还为每种物体类别提供示例图像，帮助标注员熟悉所有类别的外观。随后，标注者根据类别名称将这些图像验证并归入以下 4 个类别之一（结果列于括号中）：

1. 是，物体完全位于边界框内（*79.7%*）2. 是，物体部分位于边界框内（*16.3%*）3. 否，物体不在边界框内，但位于图像中（*1.6%*）

4. 物体不在图像中（*2.4%*）

对于类别 2，由于图像采集时这些物体在机器人的自我中心视角中仅部分可见，因此它们只部分出现在图像中。总体结果表明，真实值分割和边界框标注中的噪声非常少，质量很高，可用于训练和评估视觉感知模型。

### A.2. 关于对话引导的任务完成的用户研究

为展示 Alexa Arena 在实时用户交互方面的潜力，我们开展了一项关于对话引导的任务完成的用户研究。该研究以游戏的形式呈现给用户。在研究中，我们在 Amazon Echo 设备上运行 Alexa Arena。用户可以通过与设备交谈，使用口头命令控制虚拟环境中的智能体。

#### A.2.1. 实验设计

**参与者**。 我们招募了31名非专业参与者。

**指标**。 总体而言，我们考察任务表现和用户满意度。对于任务表现指标，我们采用任务成功率。对于用户满意度，我们收集用户对游戏总体满意度的反馈，以及用户是否认为该游戏易于学习、易于游玩、具有趣味性，以及是否愿意再次游玩。对于总体满意度，我们使用5点李克特量表；对于其他问题，则使用7点李克特量表。

**教程**。 在进行常规游戏中的任务之前，每位参与者都需要完成一个教程，帮助他们熟悉游戏机制。在教程中，参与者需要完成一项教程

任务：通过向智能体发出口头指令来制作一碗麦片。对于最初几项要求执行的动作，例如拿起麦片并将麦片倒入碗中，参与者会获得口头指令，告知他们可以使用哪些话语来执行相应动作。对于之后的动作，例如将牛奶倒入碗中，参与者在犯错之前不会获得相关话语指令。

**任务**。 要求每位参与者完成三个常规任务：

1. 第一个任务要求智能体找到控制面板并将其插入激光炮。智能体还需要开启控制激光炮发射激光的红色显示器。请注意，房间内还有一台控制冷冻射线的蓝色显示器。本任务中不应使用它。

2. 第二个任务要求智能体拾起休息室里的碗，并将其放置在变色器上。之后，智能体需要按下变色器上的红色按钮，将碗变成红色。请注意，变色器上还有绿色和蓝色按钮，智能体不应按下它们来完成本任务。

3. 第三个任务要求智能体将一罐汽水放在蓝色架子上，该架子作为冷冻射线的目标。随后，机器人可以切换蓝色显示器以发射冷冻射线，制成冰镇汽水。

**流程**。 进入游戏后，用户需要先完成教程任务以熟悉游戏，然后再进行常规任务。三个常规任务的顺序经过随机化，以避免顺序效应。如果遇到任何问题，用户可以选择跳过任务。游戏过程中，用户可以查看屏幕左上角的文字，了解任务的子目标（图 3）。用户还可以阅读房间里的便笺，获取完成任务的提示。便笺在小地图上显示为闪烁的绿色圆点。交互结束后，用户需要填写一份问卷以进行评估。

#### A.2.2. 结果与分析

31名参与者中，15人能够完成全部三个任务，另有15人能够完成两个任务。只有1人完成了一个任务。由此得出的平均任务成功率为81.72%。

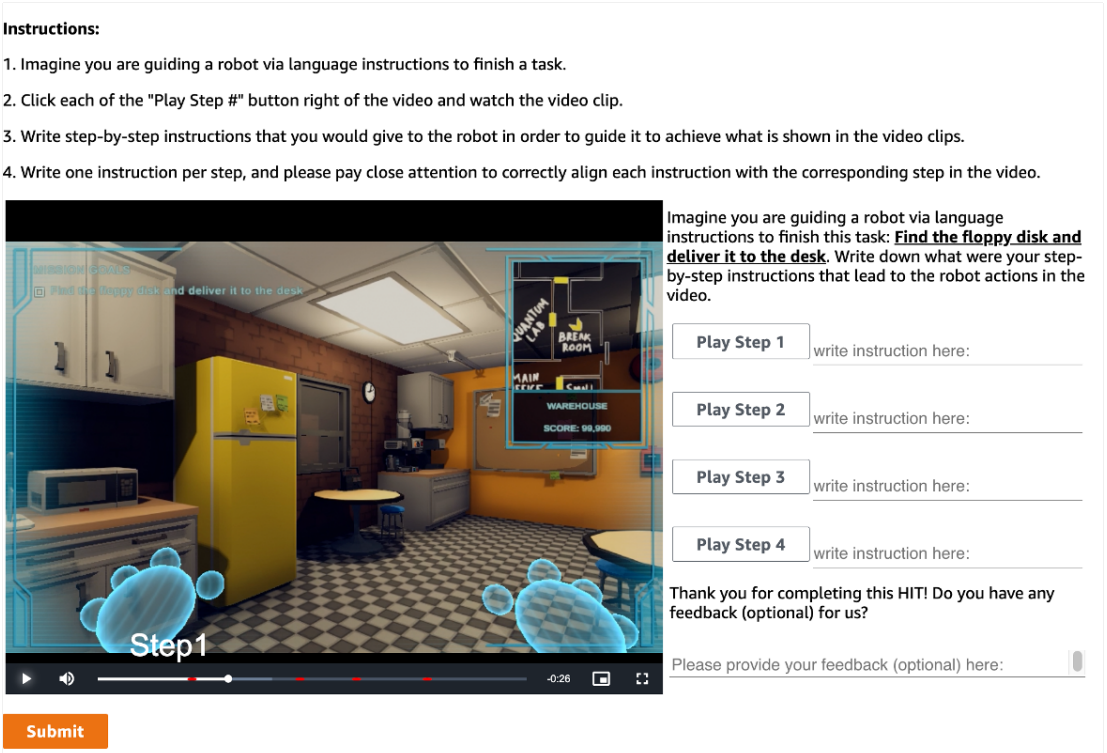
图13 展示了用户研究的结果。总体而言，大多数用户对游戏感到满意（$M = 4.03$，$SD$ =0.93）（图13a）。在游戏感知方面（图13b），大多数用户认同这款游戏易于学习（$M = 6.10$，$SD$ =0.73）、易于游玩（$M = 5.65$，$SD = 1.23$），而且玩起来很有趣（$M = 5.26$，$SD = 1.68$）。因此，77.42%的参与者表示会考虑再次玩这款游戏。

从自由形式的反馈来看，大多数参与者喜欢这款游戏。我们看到了一些这样的评论：“这是我第一次在 Alexa 上玩游戏，我真的很喜欢这款游戏。”“我觉得这款游戏很有趣，而且很容易弄清楚玩法。”“我认为这是一个不错的开端。我喜欢通过语音命令实现一些目标的想法。”关于改进，用户希望语言理解能力更强、游戏延迟更低，并在教程中获得更多指导。

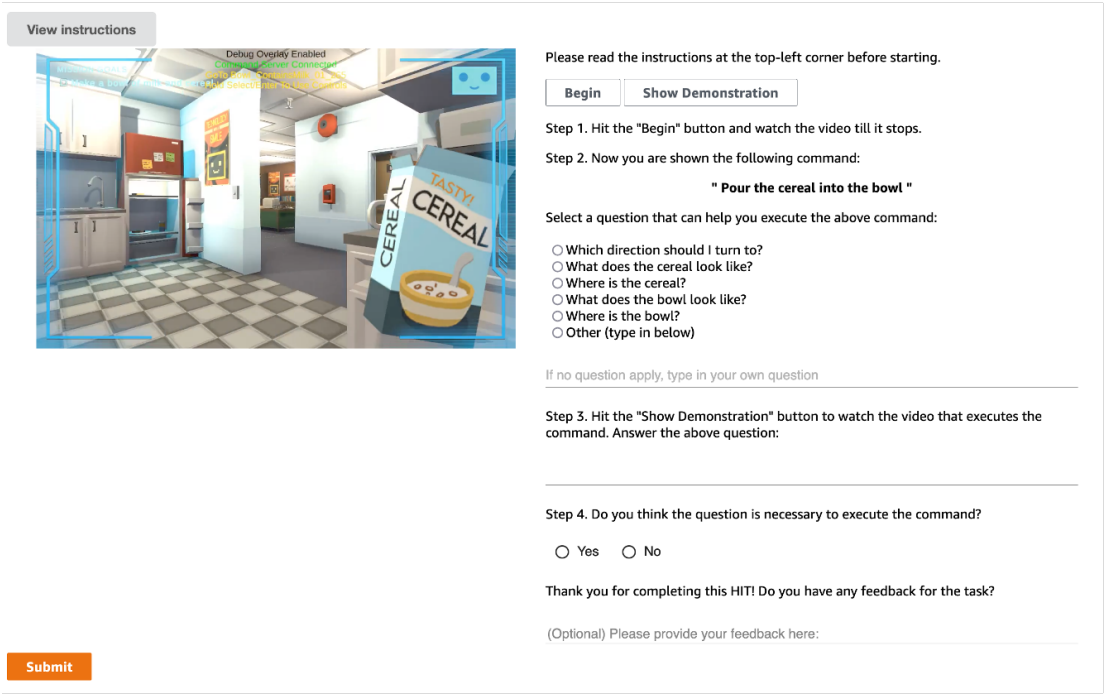
这些结果表明，Alexa Arena 平台有潜力通过将任务设计成游戏来触达普通用户，从而为大规模人机交互数据收集、系统开发以及人在环路中的评估创造可能。

## B. 布局

Alexa Arena 包含 10 个大型、可用于游戏的交互式多房间布局。每种布局都采用类似办公室的环境，并以模块化设计为特色，使房间内的物品能够重新排列和摆放。这些布局均为手工设计，在视觉上各具特色且美观有趣。办公室环境既适合融入平凡的交互元素，也能容纳奇幻的交互元素：用户在较为常见的房间里看到日常物体会感到熟悉（并且本能地知道如何与它们互动）；但进入明确标注为“实验室”的房间时，他们会期待看到未来设备和原型机器，并需要通过探索来了解其功能。得益于上述两极化的内容，我们能够设计出将富有想象力的情境与普通日常物体并置的游戏任务。Arena 中的布局规模很大：平均而言，每栋房屋都有 190 平方米的可通行区域，显著大于其他 EAI 平台中的场景（图14）。图15 展示了 Alexa Arena 中 4 栋房屋的俯视图。

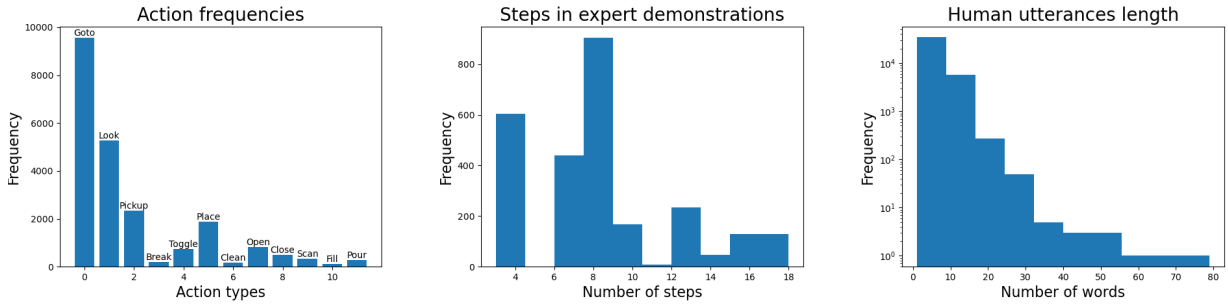



(a) 用于收集人类指令的用户界面。标注者观看专家演示，并写下指令来引导机器人。



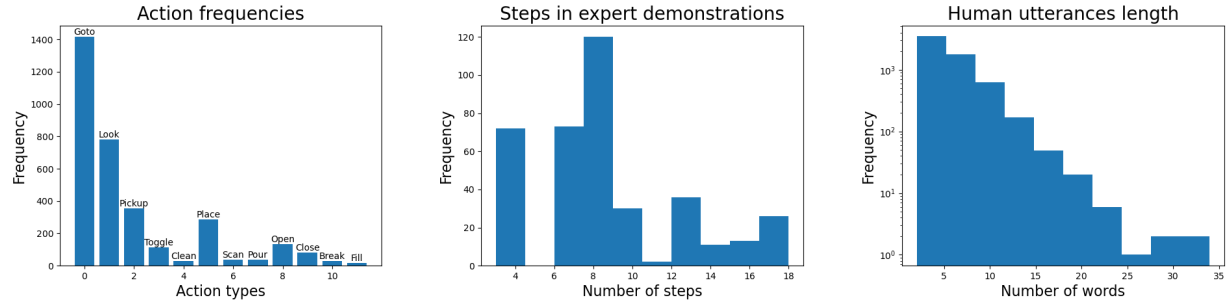

（b）用于收集问答的用户界面。标注者输入问题和答案，帮助机器人完成任务。

图 *9.* 用于数据收集的用户界面。



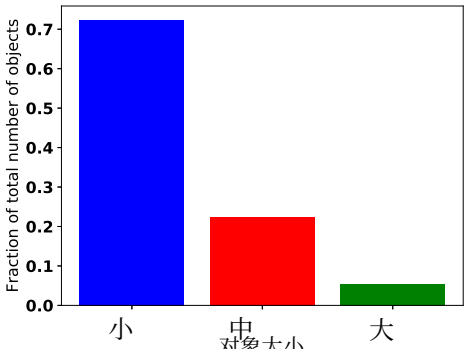

（a）动作类型的频率。 （b）专家演示中的步骤。 （c）人类话语的长度。

图 *10*。 轨迹数据集训练划分的统计数据。

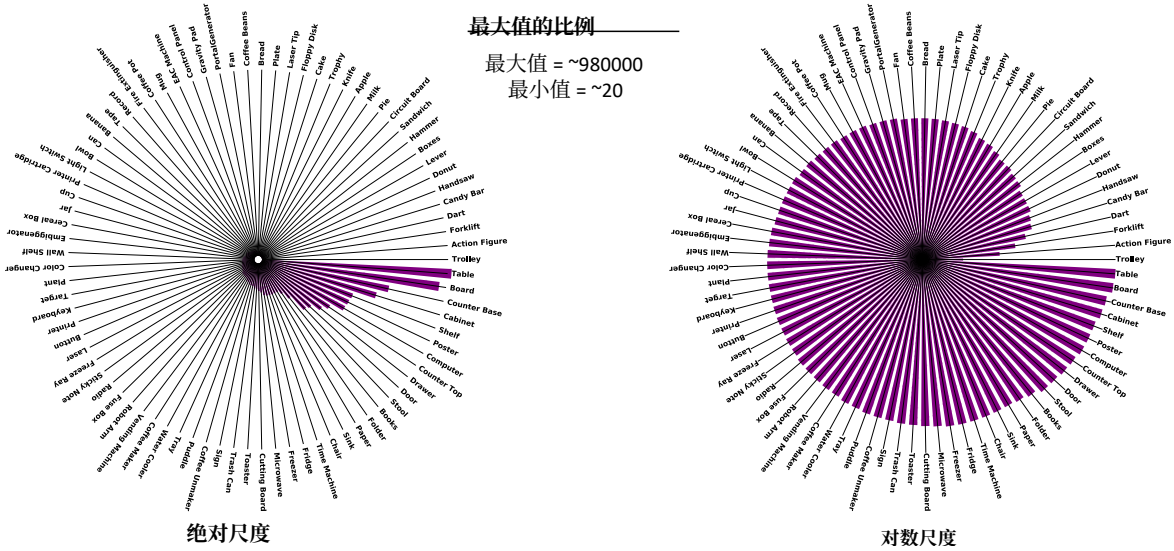



（a）动作类型的频率。 （b）专家演示中的步骤数。 （c）人类话语的长度。

图 *11*。轨迹数据集验证划分的统计信息。

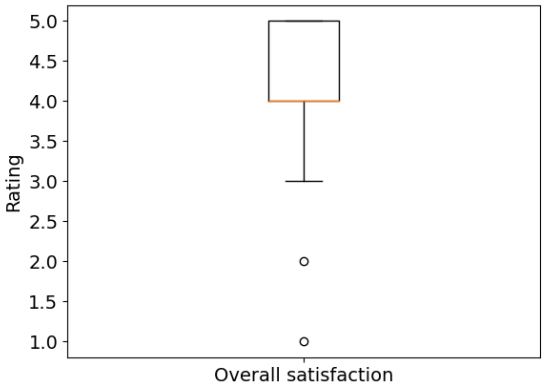



(a) 视觉数据对象大小分布。




(b) 视觉数据类别频率。

图 *12.* 视觉数据集统计




(a) 使用5点李克特量表衡量的游戏总体满意度。5表示“非常满意”，1表示“非常不满意”。橙色线表示中位数。




(b) 使用7点李克特量表衡量的用户对游戏的感知。7表示“非常同意”，1表示“非常不同意”。橙色线表示中位数。

图*13.* 用户研究结果

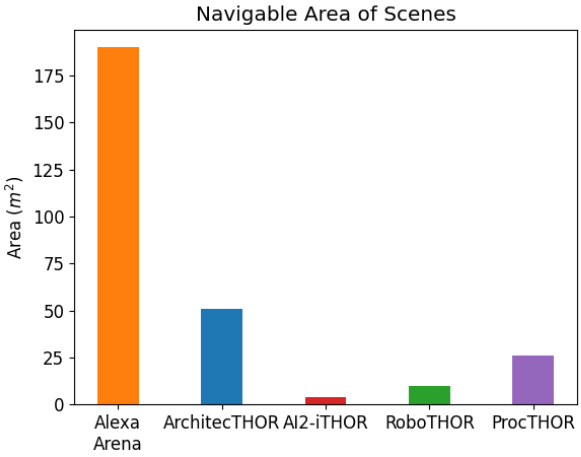



图 *14*。Alexa Arena 可导航区域的条形图，并与 ArchitecTHOR、AI2-iTHOR、RoboTHOR 和 ProcTHOR 进行比较。




图 *15*。顶部-d 4 个多房间游戏就绪布局的自有图像 在 Alexa Arena 中。