目 录

图目录

表目录

1. 发展路线：EAI 构建新概念，相关政策推动技术发展

具身智能（Embodied AI）最初是由艾伦-图灵（Alan Turing）于 1950 年提出的 “具身图灵测试”（Embodied Turing Test），旨在确定智能体是否能够展现出解决虚拟环境中问题的能力，而且能够驾驭物理世界的复杂性和不可预测性。网络空间中的智能体通常被称为非实体人工智能，而物理空间中的智能体则是实体人工智能。多模态大模型（MLMs）的最新进展为具身模型注入了强大的感知、交互和规划能力，从而开发出能与虚拟和物理环境积极交互的通用具身智能体和机器人。因此，具身智能体被广泛认为是 MLMs 的最佳载体，目前最有代表性的具身模型是 RT-2 和 RT-H。

表 1 实体人工智能和非实体人工智能

智能体种类	适应环境	物理实体	描述	代表性的智能体
非实体	网络空间	无	认知与物理实体相分离	ChatGPT, RoboGPT
实体	物理空间	机器人、汽车、其他设备	认知融入物理实体	RT-1,RT-2,RT-H

资料来源：Yang Liu《Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI》，

要让 AI 像人类一样理解这个物理世界，它必须能够以人类的方式解释和理解场景。比如，当 AI 被放路在一个房间里时，它需要能够像人类那样分析和解读周围的环境。另外，在不同领域之间建立联系，或者试图发现新知识时，传统的预编程和特定领域的专业系统已经无法满足需求。这些系统受到现有内路知识的限制，很难实现新的发现、创新和创造。让 AI 变得更聪明的关键在于利用“想象力”， 其实就是人类和其他动物依靠世界的现有模式生成的想法，它是一个非常强大的规划工具。为了让 AI 有效地规划，它需要构建一个关于世界的模型（WMs），并能够利用这个模型进行推理和决策。因此，具身认知至关重要。系统需要通过具身认知来获取知识，并进一步生成抽象的认知。

表 2 政策推动人工智能技术发展

地区	文件名称	发布时间	具体内容
上海	《上海市智能机器人标杆企业与应用场景推荐目录》	2023 年 3 月	各区产业主管部门支持推动以机器人为代表的智能终端产业发展，培育一流营商发展环境。力争到 2025 年，上海市将打造 10 家行业一流的机器人头部品牌、100 个标杆示范的机器人应用场景、1000 亿元机器人关联产业规模。
上海	《上海市推动制造业高质量发展三年行动计划（2023-2025 年）》	2023 年 6 月	瞄准人工智能技术前沿，构建通用大模型，面向垂直领域发展产业生态，建设国际算法创新基地，加快人形机器人创新发展。
北京	《北京市机器人产业创新发展行动方案 2023-2025 年》	2023 年 6 月	加紧布局人形机器人，对标国际领先人形机器人产品，支持企业和高校院所开展人形机器人整机产品、关键零部件攻关和工程化，加快建设北京市人形机器人产业创新中心。以人形机器人小批量生产和应用为目标，打造通用智能底层软件及接口、通用硬件开发配套设施等基础条件，集中突破人形机器人通用原型机和通用人工智能大模型等关键技术。
北京	《北京市促进机器人产业创新发展的若干措施》	2023 年 8 月	由机器人骨干企业牵头，整合国内外一流创新资源，组建人形机器人创新中心， 开展关键共性技术研究。支持机器人企业与“智能机器人与系统高精尖创新中心”联合开展产业化攻关。

资料来源：上海市经济和信息化委员会，上海市人民政府，北京市人民政府，

相关政策已落地，带动具身智能行业发展。例如上海市的政策重点是推动智能机器人和智能制造业的发展，目标是通过营商环境的优化和创新基地的建设，到 2025

年实现行业标杆企业和应用场景的建立。北京市的政策则侧重于机器人产业的创新发展，特

别是对高端机器人产品和国际化布局的支持，旨在推动产业生态系统的完善和技术创新。

图1 中美机器人企业落地进度一览

企业	产品型号	开始进厂时间	合作厂商	工作内容	落地进度
特斯拉	Optimus Gen2	2024.05	特斯拉	分拣电池	目前有2台Optimus人形机器人在工厂训练，预计明年超过1000台
Figure AI	Figure 01	2024.07	宝马	简单抓取	适配过程会持续12-24个月左右
优必选	Walker S	2024.01	蔚来、东风柳汽、一汽-大众青岛分公司	门锁质检、车灯盖板检测、安全带检测、贴车标、螺栓拧紧、零件安装、零件转运等	进入多家车厂实训，2024年底小规模交付
智元机器人	远征A1	暂无	均普智能、临港集团	/	/
宇树科技	H1	暂无	/	/	/
银河通用机器人	Galbot	暂无	/	/	/
达闼机器人	XR4	暂无	/	/	/
乐聚机器人	夸父	暂无	蔚来	计划在工厂检测验证	/
Apptronik	Apollo	2024.03	奔驰、GXO	汽车搬运、装配、物流配送等	奔驰在位于匈牙利的一家工厂试用数量不详的Apollo 机器人
小米机器人	CyberOne	暂无	小米汽车	计划融入小米制造、智能制造多个场景	/
波士顿动力	Atlas	暂无	现代汽车	计划在现代汽车的制造产线上应用测试	/
星动纪元	XBot-L	暂无	/	/	/
戴盟机器人	Sparky 1	暂无	/	焊接电路板	/
Agility Robotics	Digit	2023.04	亚马逊	物流	/
开普勒	先行者系列	暂无	/	/	/
Sanctuary AI	Phoenix	2024.04	麦格纳	计划将产品用于部署在麦格纳的制造业务中	/

资料来源：硬氪分析，锦囊专家，

美国在机器人工业应用和商业化方面的进展更为成熟，中国仍在探索阶段。美国和中国的机器人技术进展和落地进度存在一些显著差异。美国的机器人企业，如特斯拉和Figure AI，已在 2024 年中旬进入工厂，负责分拣电池和简单抓取的任务。Agility Robotics 与亚马逊的合作已经在 2023 年展开，推进了物流和自动化领域的实际应用。Sanctuary AI 也计划在 2024 年开始商业运营，重点放在智慧城市和建筑领域的智能服务上。相比之下，中国的机器人企业虽然在多个领域（如安防、教育和娱乐）都有布局，但整体落地进度稍慢。例如，优必选的Walker S 预计在 2024 年初开始量产，主要用于门锁质检和汽车制造领域。其他企业如小米机器人和波士顿动力的项目仍在研发阶段，逐步优化视觉和环境交互技术。

2. 技术背景：从模拟、感知、交互三方面训练 EAI

实现通用人工智能（AGI）的关键基础在于具身智能的发展。具身智能体与仅限虚拟对话的智能体（如 ChatGPT）不同，它们可以通过控制物理实体在现实和模拟环境中进行交互。该技术涵盖了多个领域，包括计算机视觉、自然语言处理和机器人技术，特别是在具身感知、具身交互以及从模拟到现实的机器人控制方面展现了显著优势。具身智能体依托于多模态大模型（MLMs）和世界模型（WMs），像“脑”一样理解虚拟与物理环境，主动感知多模态元素，并根据人类的意图进行任务分解与执行。它们不仅能够与人类互动，还能够借助知识库和工具完成复杂任务，展现出比传统深度强化学习更高的灵活性和通用性。

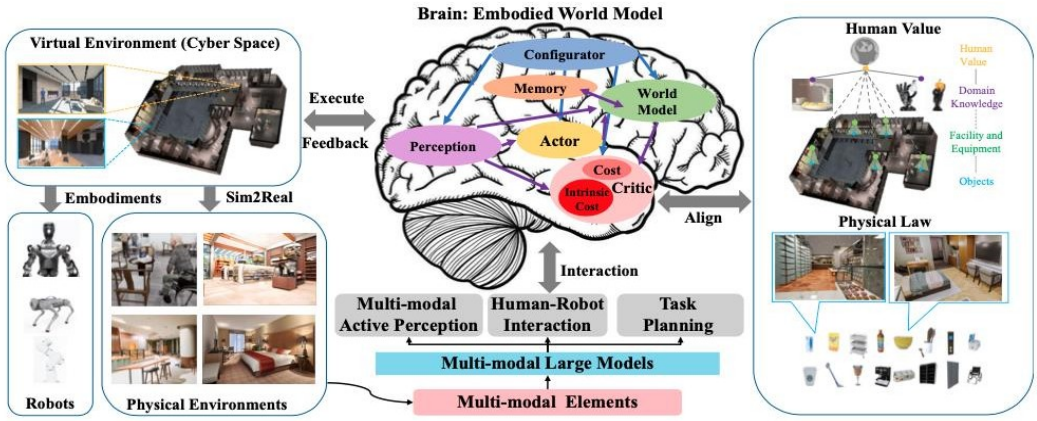
模拟器在训练具身智能中扮演了重要角色，通过提供逼真的虚拟环境进行算法开发和模型训练，帮助研究人员在降低成本、提高安全性和加速迭代的同时，将研究成果更快地转化为现实应用。同时，具身感知让智能体理解物理世界中的视觉推理和空间，技术包括视觉同步定位与绘图（vSLAM）和 3D 视觉定位，帮助智能体在动态环境中移动和互动。具身交互则强调智能体在物理或模拟空间中与环境和人类的互动能力，典型任务如具身问题解答（EQA），需要智能体主动探索环境，整合信息并执行目标导向的动作。

2.1 EAI 概念解析 · 虚拟与物理环境的结合

具身智能的发展是实现通用人工智能（AGI）的重要基础。与类似 ChatGPT 的对话智能体不同，具身智能可以通过控制物理实体并在模拟和物理环境中进行交互来实现。

目前，具身智能涵盖了计算机视觉（CV）、自然语言处理（NLP）和机器人技术等多个关键技术领域，最具代表性的是具身感知、具身交互、具身智能体和模拟到现实的机器人控制。

图2 基于 MLM 和 WM 的具身智能框架



资料来源：Yang Liu《Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI》，

基于多模态大模型（MLMs）和世界模型（WMs），具身智能体能够用“脑”来理解虚拟和物理环境，并主动感知多模态元素。主要的内容包括：1）**具身世界模型（Embodied World Model）**：这是具身智能体的核心，类似于智能体的大脑。它能够理解虚拟与物理环境，并进行多模态感知。2）**多模态感知**：智能体能够全面感知来自虚拟和物理环境的多模态元素，这意味着它可以处理不同来源的数据，如视觉、触觉等。3）**人类意图理解与对齐**：具身智能体能够充分理解人类通过语言表达的意图，并与人类的价值观对齐，确保其行为符合人类的期望。4）**任务分解与执行**：智能体具备将复杂任务分解为子任务的能力，并且能够精确执行这些任务。此外，它还能够与人类进行互动，并利用知识库和工具来完成任务。通过这些特性，具身智能体可以在复杂的动态环境中与人类自然互动，并可靠地完成任务，展现出比传统深度强化学习方法更高的灵活性和通用性。

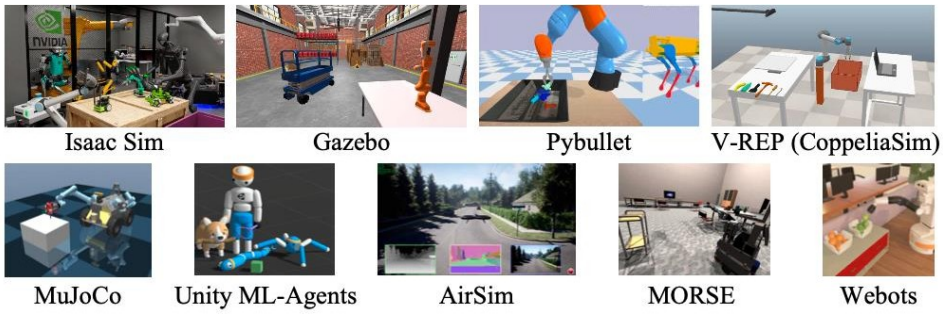
2.2 具身模拟器（Embodied Simulator）

模拟器显著提升了 AI 训练的效率，并节省了大量成本。数据匮乏一直是具身人工智能研究面临的挑战，收集真实世界的机器人数据需要花费大量时间和成本。首先，现实世界中的机器人训练需要搭建专门的物理场所，导致训练进展缓慢，效率难以提升。另外，搭建专属场地、频繁的数据收集、聘请机器人专家操作等涉及的成本很高。此外，最重要的挑战在于可重复性，因为机器人的硬件配臵、控制方法和实施框架存在巨大差异，阻碍了数据的复用性。在这种情况下，模拟器为具身人工智能的数据收集和训练提供了一种全新的解决方案。

具身模拟器对于 EAI 技术至关重要，因为它们能提供一个经济有效、可扩展且安全的实验平台。通过模拟潜在的危险场景，可以在不同环境中进行测试，支持更快的机器人原型设计，并向更广泛的研究群体开放。具身模拟器还能提供用于精确研究的受控环境，生成用于培训和评估的数据，并提供一个标准化准则。为了让具身智能体与环境互动，构建一个符合物理理论的模拟环境也十分重要，这就要求对环境的物理特性、物体的属性及其相互作用进行全面考量。

2.2.1 通用模拟器（General Simulator）

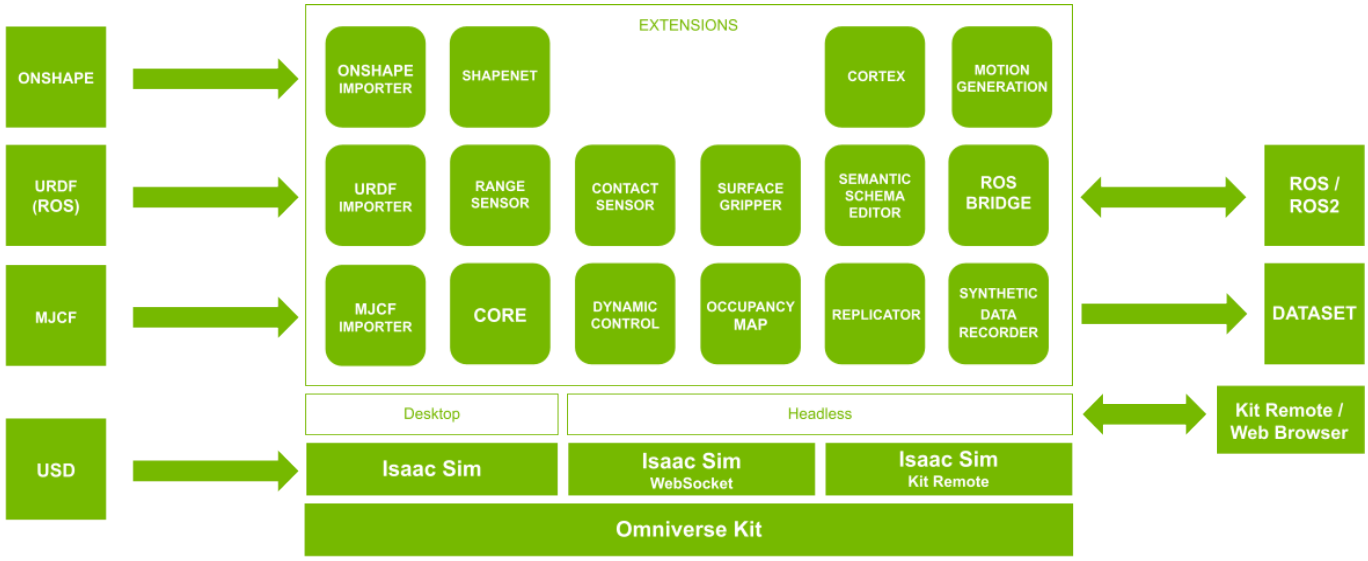
图3 通用模拟器的例子



资料来源：Yang Liu《Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI》，

具身人工智能的最终目标是将虚拟环境中的研究成果转化为现实世界中的应用。研究人员可以选择最适合自己需要的模拟器来辅助研究。通用模拟器提供了一个近似物理世界的虚拟环境，可以进行算法开发和模型训练，在成本、时间和安全性方面都有显著优势。

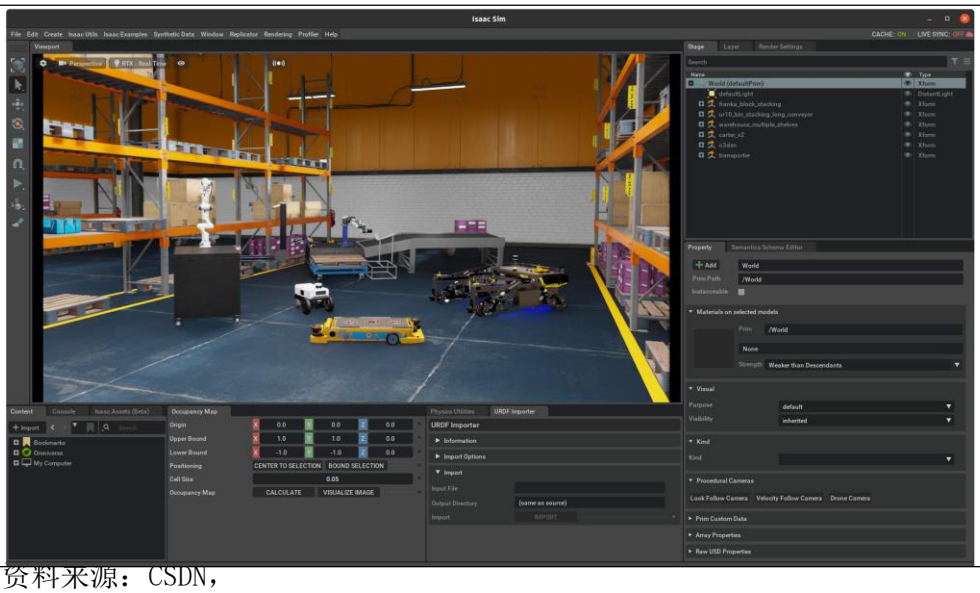
图 4 Isaac Sim 架 构



资料来源：Isaac Sim Documentation，CSDN，

Isaac Sim 是一个专为 NVIDIA Omniverse 平台开发的机器人仿真工具包，它提供了构建仿真机器人世界和进行实验所需的大部分功能。Isaac Sim 可以接受来自不同来源的输入，比如 Onshape、URDF、MJCF、USD，其中 USD 直接导入 Isaac Sim，其他类型的输入则会通过对应的 importer 插件进行导入。Onshape 是一种基于云的计算机辅助设计（CAD）软件，用于进行三维建模和设计工作。类似于 Fusion360。URDF（Unified Robot Description Format）是一种 XML 文件格式，用于描述机器人模型的几何形状、连接性、关节、传感器和其他相关信息。在这个架构中，USD（Universal Scene Description）用作场景描述，用于在不同工具之间进行内容创建和交换。目前 USD 正在广泛应用，不仅在视觉效果社区，还在建筑、设计、机器人技术、制造和其他领域中得到采用。

图5 Isaac Sim 工作界面



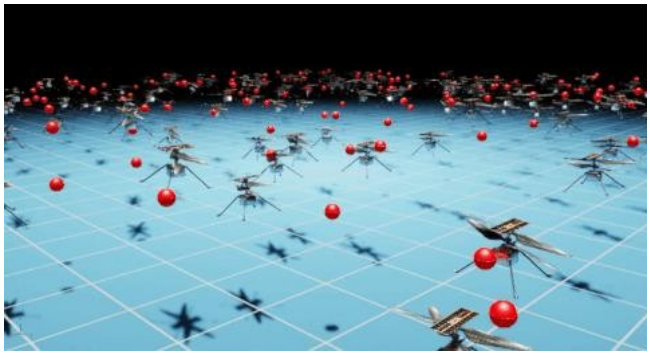
资料来源：CSDN，

该工具包还提供了创建稳健、物理精确的仿真和合成数据集所需的工具和工作流程。Isaac Sim 支持常见的机器人框架，如 ROS/ROS2，允许用户通过这些框架进行导航和操作应用。此外，Isaac Sim 能够模拟来自多种传感器的数据，包括 RGB-D、激光雷达和 IMU，适用于各种计算机视觉技术，如域随机化、地面真值标注、分割和边界框的生成。

图6 Isaac 模拟机械手臂

资料来源：IsaacGymEnvs，CSDN，

图7 Isaac 模拟无人机飞行



资料来源：IsaacGymEnvs，CSDN，

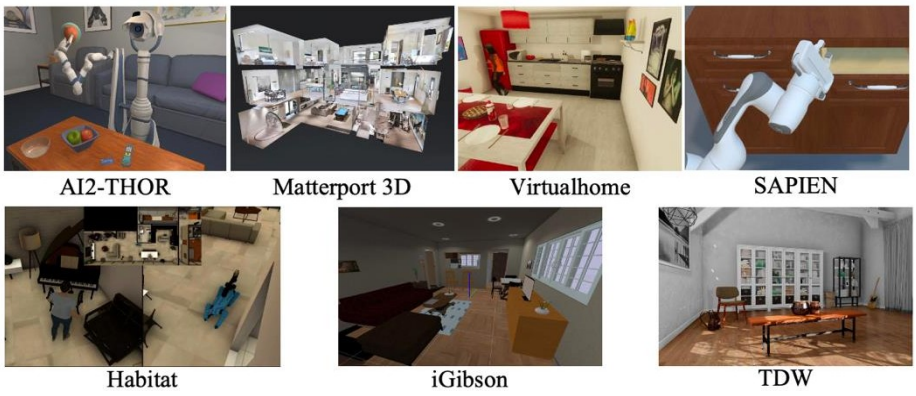
机器人仿真是利用计算机技术来模拟机器人运动、感知和互动的过程。这一过程涵盖了机器人硬件和软件系统的模拟，以便在虚拟环境中对机器人的算法和控制策略进行测试、开发和验证。其中的好处包括 **1）成本控制**：仿真环境能显著降低机器人开发与测试成本，避免了对昂贵硬件和设备的依赖。若不使用仿真，而采用大量真实机器人进行测试，将面临硬件组装、调试及损坏等高昂的时间与经济成本。尤其对于特殊测试场景，如沙漠或核电站，搭建逼真测试环境的成本同样极高。**2）安全性保障**：例如在工业机械臂、无人机等机器人设备调试中，无人机失控坠毁或机械臂故障会对企业的人员安全构成威胁。**3）快速迭代**：仿真环境避免了对真实机器人的繁琐调试，例如为一千台机器人重新烧录固件或修改搭载的算法，从而节省了大量的调试时间。并且模拟器允许开发人员快速迭代机器人的算法和控制器，以优化性能和功能。

2.2.2 基于真实世界的模拟器（Real-Scene Based Simulators）

在室内活动中实现通用具身智能一直是 AI 研究领域的重点。这些具身智能体需要深入理解人类的日常生活，并执行复杂的具身任务，如室内环境中的导航和交互。为了满

足这些复杂任务的需求，模拟环境需要尽可能接近真实世界，这就对模拟器的复杂性和逼真度提出了很高的要求。因此，基于真实世界环境的模拟器应运而生。这些模拟器大多从现实世界收集数据，创建逼真的三维资产，并使用 UE5（虚幻 5）和 Unity 等三维游戏引擎构建场景。丰富而逼真的场景使基于真实世界环境的模拟器成为研究家居活动中的体现式人工智能的首选。

图8 基于真实世界的模拟器实例

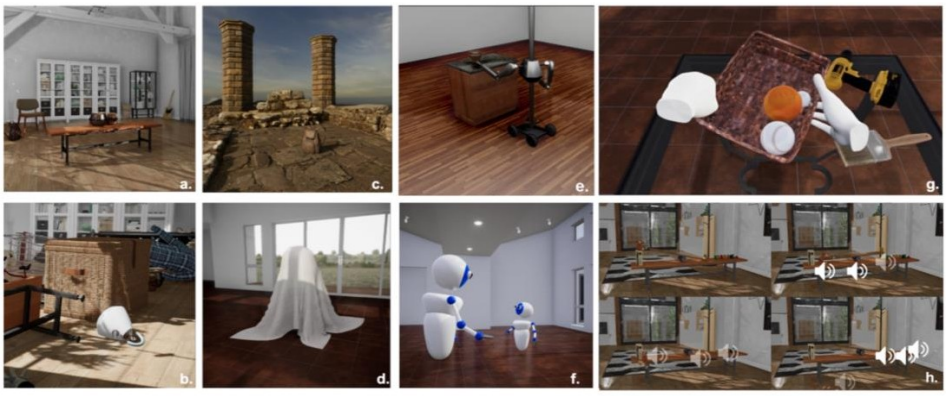


资料来源：Yang Liu《Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI》，

在腾讯网援引映维网的文章中指出，2021 年，麻省理工学院（MIT）、MIT-IBM 沃森人工智能实验室、哈佛大学和斯坦福大学的研究人员开发了一个名为 ThreeDWorld（TDW）的平台，并希望创造一个类似于《黑客帝国》的丰富虚拟世界。TDW 能够模拟室内和室外的高保真音频和视频环境，并允许用户像在现实生活中一样根据物理定律与对象进行交互。当发生相互作用时，系统能够计算并执行流体、柔体和刚体的对象方向、物理特征和速度，从而产生精确的碰撞和撞击声音。

TDW 支持在三维环境中模拟移动智能体和对象之间的高保真感觉数据和物理交互。独特的特性包括：实时接近照片真实感的图像渲染；各种物质类型的真实物理交互作用，包括布、液体和可变形物体；具身智能体的可定制“智能体”； 并支持人类与 VR 设备的交互。 TDW 的 API 允许多个智能体在模拟中交互，并返回代表世界状态的传感器和物理数据范围。 Yang Liu 等人介绍了 TDW 在计算机视觉、机器学习和认知科学等新兴研究方向上的初步实验，包括多模态物理场景理解、物理动力学预测、多智能体交互、“像孩子一样学习”的模型，以及人类和神经网络的注意力研究。

图9 ThreeDWorld（TDW）设计展示



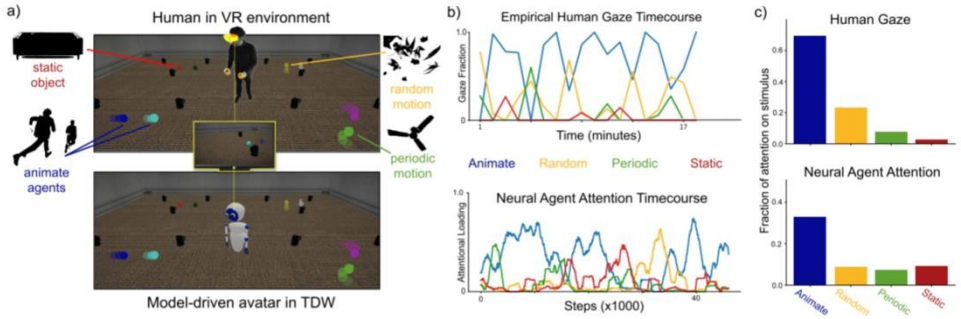
资料来源：Yang Liu《ThreeDWorld: A Platform forInteractive Multi-Modal Physical Simulation》，CSDN，

利用 TDW 的多智能体 API 的灵活性，可以创建各种多智能体交互设路的实现。 其中包括一个“观察者”智能体被安路在一个有多个无生命物体的房间里，与几个不同控

制的“行动者”智能体一起（图 9a)。“行动者”智能体由硬编码或交互策略控制，实现

对象操作、追逐和隐藏以及运动模仿等行为。在这种情况下，人类观察者只被要求看他们想看的任何东西，而虚拟观察者寻求最大限度地提高其预测同一显示中参与者行为的能力，根据“进展好奇心”的度量来分配其注意力，该度量寻求估计哪些观察最有可能增加观察者做出参与者预测的能力。

图10 多智能体互动和 VR 能力



资料来源：CSDN，

2.3 具身感知（Embodied Preception）

具身感知未来主要的发展方向是以智能体为中心的视觉推理。与仅仅识别图像中的物体不同，具有具身感知能力的智能体必须在物理世界中移动并与环境互动。这就要求对三维空间和动态环境有更深入的了解。

2.3.1 视觉同步定位和绘图 (vSLAM)

SLAM (Simultaneous Localization And Mapping，同步定位与地图构建)，主要为了解决移动机器人在未知环境运行时定位导航与地图构建的问题。SLAM 能够解决机器人在陌生环境中的定位、环境感知、移动方向等问题。机器人可以配臵多种传感器来实现 SLAM，包括激光雷达（3D，2D），毫米波雷达，超声波，RGB-D，摄像头（单目，多目）等，通常根据使用场景、制造成本、设备功率、算力的需求与约束，机器人采用不同传感器或组合的解决方案，以减少误差并提高准确性。目前两个主流的解决方案是基于激光雷达的 Lidar SLAM 以及基于摄像头的 Visual SLAM。

VSLAM 即 Visual Simultaneous Localization and Mapping，主要是指如何用相机解决定位和建图问题。当用相机作为传感器时，通过一张张连续运动的图像(它们形成一段视频)，从中推断相机的运动，以及周围环境的情况。VSLAM 的技术框架主要由 5 部分组成，包括传感器数据预处理、前端、后端、回环检测、建图。前端，又称为视觉里程计（visual odometry，简称 VO），主要是研究如何根据相邻帧图像定量估算帧间相机的运动。通过把相邻帧的运动轨迹串起来，就构成相机载体（如机器人）的运动轨迹， 解决定位的问题，然后根据估算的每个时刻相机的位臵，计算出各像素的空间点的位臵， 就得到地图。