

# A Survey of Embodied AI: From Simulators to Research Tasks

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, Cheston Tan

<https://arxiv.org/abs/2103.04918>

## 摘要

从“互联网人工智能”时代到“具体化人工智能”时代，人工智能算法和代理不再主要来自互联网的图像、视频或文本数据集中学习，这是一个新兴的范式转变。相反，它们通过与环境的互动，从一种类似于人类的以自我为中心的感知中学习。因此，对嵌入式人工智能模拟器的需求大幅增长，以支持各种嵌入式人工智能研究任务。对嵌入式人工智能日益增长的兴趣有利于对通用人工智能(AGI)的更大追求，但目前还没有对这一领域进行当代和全面的调查。本文旨在为**具身智能**领域提供一个百科全书式的概述，从其模拟器到其研究。通过评估我们提出的七个特征的九个当前嵌入AI模拟器，本文旨在了解**模拟器在嵌入AI研究中的规定及其局限性**。最后，本文综述了嵌入人工智能的**三个主要研究任务——视觉探索、视觉导航和嵌入问答(QA)**，涵盖了最先进的方法、评估指标和数据集。最后，通过调查该领域揭示的新见解，本文将为任务模拟器的选择提供建议，并为该领域的未来方向提出建议。

## Introduction

Embodied AI is the belief that true intelligence can emerge from the interactions of an agent with its environment.

但就目前而言，具身人工智能是将视觉、语言和推理等传统智能概念整合到人工化身中，以帮助解决虚拟环境中的人工智能问题。

评估具身导航的综述论文: On Evaluation of Embodied Navigation Agents, 2018(<https://arxiv.org/abs/1807.06757>)

This paper covers the following nine embodied AI simulators that were developed over the past four years: **DeepMind Lab** [12], **AI2-THOR** [13], **CHALET** [14], **VirtualHome** [15], **VRKitchen** [16], **Habitat-Sim** [17], **iGibson** [18], **SAPIEN** [19], and **ThreeDWorld** [20]. (所选择的模拟器是为通用智能任务设计的，不像游戏模拟器，它只用于训练强化学习代理。这些嵌入的AI模拟器在计算机模拟中提供真实世界的逼真表示，主要采用房间或公寓的配置，为环境提供某种形式的约束。这些模拟器中的大多数至少包含一个物理引擎、Python API和可以在环境中控制或操纵的人工代理。)

**DeepMind Lab**: <https://deepmind.google/discover/blog/open-sourcing-deepmind-lab/>

**AI2-THOR**: AI2-THOR: An Interactive 3D Environment for Visual AI、<https://arxiv.org/abs/1712.05474>

**CHALET**: CHALET: Cornell House Agent Learning Environment、<https://arxiv.org/abs/1801.07357>

**VirtualHome**: VirtualHome: Simulating Household Activities via Programs、<https://arxiv.org/abs/1806.07011>

**VRKitchen**: VRKitchen: an Interactive 3D Virtual Environment for Task-oriented Learning、<https://arxiv.org/abs/1903.05757>

**Habitat-Sim**: Habitat: A Platform for Embodied AI Research、<https://arxiv.org/abs/1904.01201>

**iGibson** : Interactive Gibson Benchmark (iGibson 0.5): A Benchmark for Interactive Navigation in Cluttered Environments、<https://arxiv.org/abs/1910.14442>

**SAPIEN** : SAPIEN: A SimulATED Part-based Interactive ENvironment、<https://arxiv.org/abs/2003.08515>

**ThreeDWorld**: ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation、<https://arxiv.org/abs/2007.04954>

Focus on:

**visual exploration (视觉探索)** , **visual navigation (视觉导航)** and **embodied QA (嵌入式QA)** : (视觉探索是视觉导航中非常有用的组件，用于现实情境，而具体化的QA则进一步涉及建立在视觉和语言导航之上的复杂QA功能。)

# SIMULATORS FOR EMBODIED AI

## A:Embodied AI Simulators

各Simulator对比

TABLE I  
SUMMARY OF EMBODIED AI SIMULATORS. ENVIRONMENT: GAME-BASED SCENE CONSTRUCTION (G) AND WORLD-BASED SCENE CONSTRUCTION (W). PHYSICS: BASIC PHYSICS FEATURES (B) AND ADVANCED PHYSICS FEATURES (A). OBJECT TYPE: DATASET DRIVEN ENVIRONMENTS (D) AND OBJECT ASSETS DRIVEN ENVIRONMENTS (O). OBJECT PROPERTY: INTERACT-ABLE OBJECTS (I) AND MULTI-STATE OBJECTS (M). CONTROLLER: DIRECT PYTHON API CONTROLLER (P), VIRTUAL ROBOT CONTROLLER(R) AND VIRTUAL REALITY CONTROLLER (V). ACTION: NAVIGATION (N), ATOMIC ACTION (A) AND HUMAN-COMPUTER INTERACTION (H). MULTI-AGENT: AVATAR-BASED (AT) AND USER-BASED (U). THE SEVEN FEATURES CAN BE FURTHER GROUPED UNDER THREE SECONDARY EVALUATION FEATURES; REALISM, SCALABILITY AND INTERACTIVITY.

Year	Embodied AI Simulator	Environment (Realism)	Physics (Realism)	Object Type (Scalability)	Object Property (Interactivity)	Controller (Interactivity)	Action (Interactivity)	Multi-agent (Interactivity)
2016	DeepMind Lab	G	-	-	-	P, R	N	-
2017	AI2-THOR	G	B	O	I, M	P, R	A, N	U
2018	CHALET	G	B	O	I, M	P	A, N	-
2018	VirtualHome	G	-	O	I, M	R	A, N	-
2019	VRKitchen	G	B	O	I, M	P, V	A, N, H	-
2019	Habitat-Sim	W	-	D	-	-	N	-
2019	iGibson	W	B	D	I	P, R	A, N	U
2020	SAPIEN	G	B	D	I, M	P, R	A, N	-
2020	ThreeDWorld	G	B, A	O	I	P, R, V	A, N, H	AT

TABLE II  
COMPARISON OF EMBODIED AI SIMULATORS IN TERMS OF ENVIRONMENT CONFIGURATION, SIMULATION ENGINE, TECHNICAL SPECIFICATION, AND RENDERING PERFORMANCE.

Embodied AI Simulator	Environment Configuration	Simulation Engine	Technical Specification	Rendering Performance
DeepMind Lab	Customized environment	Quake II Arena Engine	6-core Intel Xeon CPU and an NVIDIA Quadro K600 GPU	158 fps/thread
AI2-THOR	120 rooms, 4 categories	Unity 3D Engine	Intel(R) Xeon(R) CPU E5-2620 v4 and NVIDIA Titan X	240 fps/thread
CHALET	58 rooms, 10 houses	Unity 3D Engine	-	-
VirtualHome	6 apartments with multiple jointed rooms	Unity 3D Engine	-	Customized frame rate
VRKitchen	16 kitchens	Unreal Engine 4	Intel(R) Core(TM) i7-7700K processor and NVIDIA Titan X	15 fps/thread
Habitat-Sim	Multiple datasets	-	Xeon E5-2690 v4 CPU and Nvidia Titan Xp GPU	10,000 fps/thread
iGibson	Gibson V1	-	Modern GPU	1000 fps/thread
SAPIEN	Customized environment	PhysX Physical engine and ROS	Intel i7-8750 CPU and an Nvidia GeForce RTX 2070 GPU	700 fps/thread
ThreeDWorld	Customized environment	Unity 3D Engine	Intel i7-7700K GPU: NVIDIA GeForce GTX 1080	168 fps/thread

## 七项评估特征

环境(Environment):

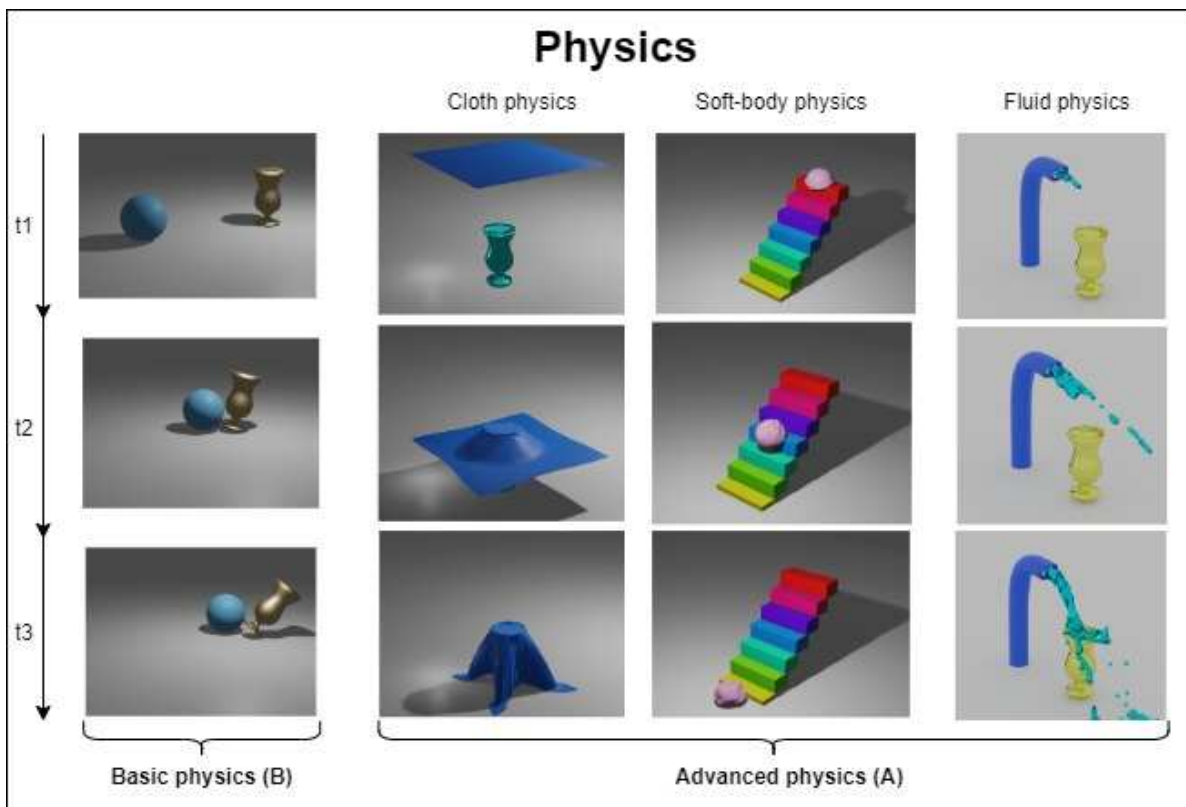


game-based scene construction (G) and world-based scene construction (W).

G:3D建模得到的。

W:基于世界的场景是通过对物体和环境的真实扫描来构建的。

## 物理学(Physics):



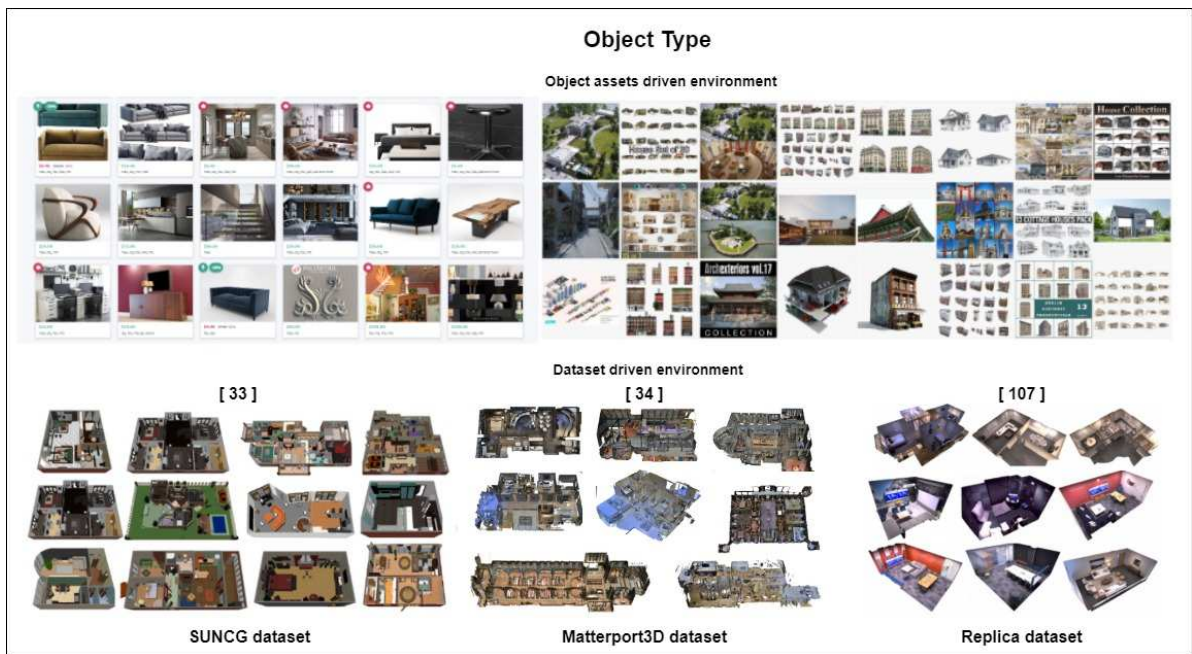
构建真实的环境、在代理和对象之间或对象和对象之间构建真实的交互，以模拟真实世界的物理属性。

basic physics features (B) and advanced physics features (A)

B:碰撞(collision)、刚体动力学(rigid-body dynamics)、重力建模(gravity modelling)

A:布(cloth)、流体(fluid)、软体物理(soft-body physics)

## 物体类型(objectType):



数据集驱动环境：其中的物体主要来自现有的物体数据集。

SUNCG、Matterport3D、Cibson等

资产驱动环境：其中的物体来自网络

Unity 3D、游戏资产

然而，在资产驱动对象中，要保证3D对象模型的质量比在数据集驱动对象中更难。根据我们的回顾，基于游戏的嵌入AI模拟器更有可能从资产存储中获取它们的对象数据集，而基于世界的模拟器则倾向于从现有的3D对象数据集中导入它们的对象数据集。

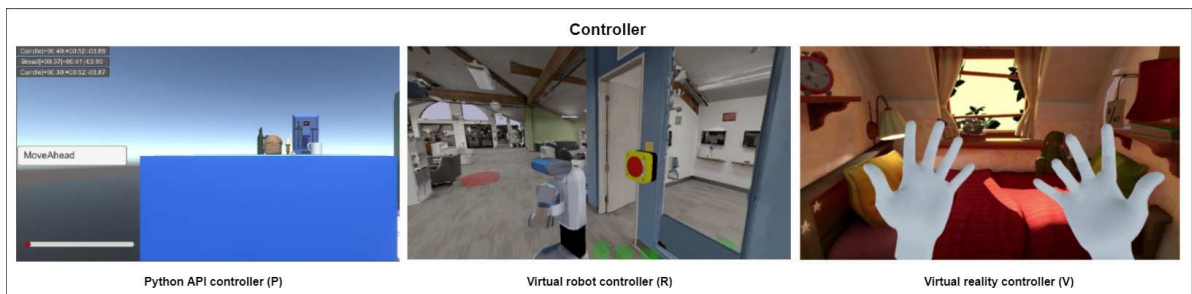
## 物体属性(Object Property):

一些模拟器只启用具有碰撞等基本交互性的对象。高级模拟器使对象具有更细粒度的交互性，例如多状态更改。

具有可交互物体(interact-able objects, I)

多状态物体(multiple-state objects, M)

## 控制器(Controller):



用户与模拟器之间有不同类型的控制器接口，从直接 Python API控制器(P)，虚拟机器人控制器(R)到虚拟现实控制器(V)。具身机器人允许与现有的现实世界机器人(如 Universal Robot5(UR5)和TurtleBot V2)进行虚拟交互，并可直接使用 ROS 界面进行控制。

## 行动(Action):

本文将它们分为机器人操作的三个层次:导航(N)、原子动作(A)和人机交互(H)。

导航(Navigation):

导航是最底层,是所有具身人工智能模拟器的共同特征[38]。它由智能体在虚拟环境中的导航能力定义。

原子动作(atomicaction, A):

原子动作作为人工智能智能体提供了对感兴趣的物体进行基本离散操作的手段,在大多数人工智能模拟器中都能找到。

人机交互(human-computerinteraction, H):

人机交互是虚拟现实控制器的成果,因为它能让人类控制虚拟智能体与模拟世界实时学习和互动。

## 多智能体(Multi-agent):

目前涉及多agent强化学习的研究很少。只有AI2-THOR, iGibson和ThreeDWorld等少数模拟器配备了多agent设置。

基于强化学习的多机器人训练:目前仍在OpenAI Gym环境中进行

第一个是ThreeDWorld中的基于化身(AT)的多代理,它允许人工代理和模拟化身之间的交互。

第二种是AI2-THOR中的基于用户(user-based, U)的多智能体,它可以扮演双重学习网络的角色,通过与仿真中的其他人工智能体交互来学习,以完成共同的任务。

## B:Comparison of Embodied AI Simulators

---

**真实性:**指的是模拟器中3D环境的逼真程度,包括环境的物理外观和物理特性的模拟。

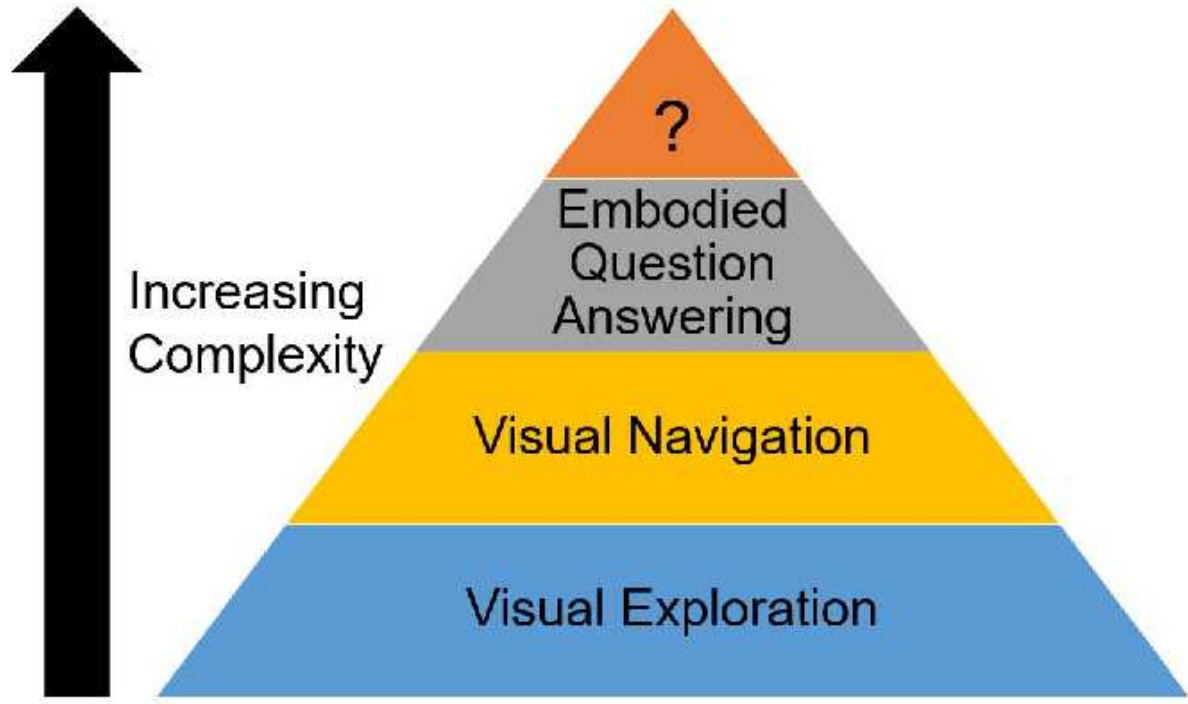
**可扩展性:**描述了模拟器中3D环境的扩展能力,可以通过采集更多的真实世界3D扫描数据或购买更多3D资产来实现。

**互动性:**涉及到物体属性、控制器、动作和多智能体的交互,这些因素直接影响模拟器中的交互体验。

## RESEARCH IN EMBODIED AI

---





具身智能研究任务的三大类型是视觉探索、视觉导航和具身QA。我们将把重点放在这三种任务上，因为大多数现有的具身人工智能论文都集中在这些任务上，或者利用为这些任务引入的模块来为更复杂的任务（如视听导航）建立模型。从探索到QA，任务的复杂性不断增加。我们将从视觉探索开始，然后是视觉导航，最后是体现式QA。如图 5 所示，这些任务中的每一个都是下一个任务的基础，形成了一个具身人工智能研究任务的金字塔结构，进一步表明了具身人工智能的自然发展方向。我们将重点介绍每个任务的重要方面，从摘要、方法、评估指标到数据集。这些任务的详细信息见表。

TABLE III  
SUMMARY OF EMBODIED AI RESEARCH TASKS. EVALUATION METRIC: AMOUNT OF TARGETS VISITED (ATV), DOWNSTREAM TASKS (D), SUCCESS WEIGHTED BY PATH LENGTH (SPL), SUCCESS RATE (SR), PATH LENGTH RATIO (PLR), ORACLE SUCCESS RATE (OSR), TRAJECTORY/EPISODE LENGTH (TL / EL), DISTANCE TO SUCCESS / NAVIGATION ERROR (DTS / NE /  $d_T$ ), GOAL PROGRESS (GP /  $d_\Delta$ ), ORACLE PATH SUCCESS RATE (OPSR), SMALLEST DISTANCE TO TARGET AT ANY POINT IN AN EPISODE ( $d_{min}$ ), PERCENTAGE OF EPISODES AGENT ENDS NAVIGATION FOR ANSWERING BEFORE MAX EPISODE LENGTH ( $\%stop$ ), PERCENTAGE OF QUESTIONS AGENT TERMINATES IN THE ROOM CONTAINING THE TARGET OBJECT ( $\%r_T$ ), PERCENTAGE OF QUESTIONS WHERE THE AGENT ENTERS THE ROOM CONTAINING THE TARGET OBJECT AT LEAST ONCE ( $\%r_e$ ), INTERSECTION OVER UNION FOR TARGET OBJECT (IoU), HIT ACCURACY BASED ON IoU ( $h_T$ ), MEAN RANK OF THE GROUND-TRUTH ANSWER IN QA PREDICTIONS (MR) AND QA ACCURACY (ACC).

Task	Method / Category	Publication	Year	Simulator	Dataset	Evaluation Metric
Visual Exploration	Curiosity	Chaplot et al. [43]	2020	Habitat-Sim	Matterport3D, Gibson V1	ATV
		Chaplot et al. [44]	2020	Habitat-Sim	Matterport3D, Gibson V1	ATV, D
	Reconstruction	Ramakrishnan et al. [45]	2020	Habitat-Sim	Matterport3D, Gibson V1	ATV, D
		Ramakrishnan et al. [46]	2020	Habitat-Sim	Matterport3D	ATV, D
Visual Navigation	Point Navigation	Narasimhan et al. [47]	2020	Habitat-Sim	Matterport3D	ATV, D
		Wijmans et al. [48]	2019	Habitat-Sim	Matterport3D, Gibson V1	SPL, SR
		Georgakis et al. [49]	2019	Habitat-Sim	Matterport3D	SR, PLR
		Ye et al. [50]	2020	Habitat-Sim	Gibson V1	SPL, SR
		Chaplot et al. [44]	2020	Habitat-Sim	Matterport3D, Gibson V1	SPL, SR
		Ramakrishnan et al. [45]	2020	Habitat-Sim	Matterport3D, Gibson V1	SPL, SR
		Ramakrishnan et al. [46]	2020	Habitat-Sim	Matterport3D	SPL
		Narasimhan et al. [47]	2020	Habitat-Sim	Matterport3D	SPL, SR
		Claudia et al. [50]	2020	iGibson	Gibson V1	SR
	Object Navigation	Wortsman et al. [51]	2019	AI2-THOR	-	SPL, SR
		Campar et al. [52]	2020	Habitat-Sim	Matterport3D	SPL, SR, DTS
		Du et al. [53]	2020	AI2-THOR	-	SPL, SR
		Chaplot et al. [54]	2020	Habitat-Sim	Matterport3D, Gibson V1	SPL, SR, DTS
		Shen et al. [55]	2020	iGibson	Gibson V1	SR
		Wahid et al. [56]	2020	-	Gibson V1	SPL, SR
		Yang et al. [57]	2020	AI2-THOR	-	SPL, SR
	Vision-and-Language Navigation	Anderson et al. [58]	2018	-	Room-40-Room	SR, OSR, TL, NE
		Zhu et al. [59]	2020	-	Room-40-Room	SPL, SR, OSR, TL, NE
		Zhu et al. [60]	2020	-	Cooperative Vision-and-Dialog Navigation	SR, OSR, GP, OPSR
Embodied Question Answering	Question Answering	Das et al. [61]	2018	-	EQA	$d_T, d_\Delta, d_{min}, \%r_T, \%r_e, \%stop, MR$
		Das et al. [62]	2018	-	EQA	$d_T, d_\Delta, Acc$
	Multi-target Question Answering	Yu et al. [63]	2019	-	MT-EQA	$d_T, d_\Delta, \%r_T, \%stop, IoU, h_T, Acc$
	Interactive Question Answering	Gordon et al. [64]	2018	AI2-THOR	IQUAD V1	EL, Acc
		Tan et al. [65]	2020	AI2-THOR	IQUAD V1	EL, Acc

## A. Visual Exploration

在传统的机器人技术中，探索是通过被动或主动的同步定位和绘图(simultaneous localisation and mapping)但纯SLAM)来完成的，以建立环境地图。然后将该地图与定位和路径规划一起用于导航任务。SLAM 的研究非常深入粹的几何方法仍有改进的余地。由于它们依赖于传感器，因此容易受到测量噪声的影响，需要进行大量的微调。另一方面，基于学习的方法通常使用 RCB和/或深度传感器，对噪声具有更强的鲁棒性。此外，视觉探索中基于学习的方法允许人工智能结合语义理解(如环境中的物体类型)，并以无监督的方式归纳其先前所见环境的知识，以帮助理解新环境。这就减少了对人类的依赖，从而提高了效率。

## (1)方法

视觉探索中的非基线方法通常被形式化为部分观察的马尔可夫决策过程 (POMDP) [77]。POMDP 可以用 7 元组  $(S, A, T, R, \Omega, O, \gamma)$  表示, 其中状态空间  $S$ 、动作空间  $A$ 、转移分布  $T$ 、奖励函数  $R$ 、观察空间  $\Omega$ 、观察分布  $O$  和折扣因子  $\gamma \in [0, 1]$ 。一般来说, 这些方法被视为 POMDP 中的特定奖励函数



## (2)评估指标

**访问的目标数量:** 考虑了不同类型的目标, 如区域[44]、[86]和有趣的对象[72]、[87]。访问的区域度量有一些变体, 例如  $m$  的绝对覆盖面积  $m^2$  以及在场景中被探索的区域的百分比。

**对下游任务的影响:** 视觉探索性能也可以通过其对视觉导航等下游任务的影响来衡量。

## (3)数据集

对于视觉探索, 一些流行的数据集包括Matterport3D和Gibson V1。Matterport3D和Gibson V1都是逼真的RGB数据集, 具有嵌入AI的有用信息, 如深度和语义分割。

## B. Visual Navigation

在视觉导航中, 智能体可在有或没有外部先验或自然语言指令的情况下, 在二维环境中向目标导航。这项任务使用了许多类型的目标, 如点, 物体, 图像和区域。本文重点讨论点和物体。它们可以进一步与感知输入和语言等规范相结合, 从而构建出更复杂的视觉导航任务, 如带有先验的导航、视觉与语言导航, 甚至是具身QA(Embodied QA)。

经典的导航方法通常由人工设计的子组件组成, 如定位, 映射, 路径规划和运动。具身人工智能中的视觉导航旨在从数据中学习这些导航系统, 以减少特定情况下的手工工程, 从而便于与数据驱动学习方法(如QA)性能优越的下游任务整合。还有一些混合方法, 旨在将两者的优点结合起来。基于学习的方法对传感器测量噪声的鲁棒性更强, 因为它们使用RGB 和/或深度传感器, 并能结合对环境的语义理解。此外, 这些方法还能让智能体在无监督的情况下, 将其对以前所见环境的了解用于帮助理解新环境, 从而减少人力。

著名的挑战赛有icibson sim2Real challenge、Habitat Challenge 和 RoboTHOR Challenge。

(1)方法

在点导航中，智能体的任务是导航到一个特定的点。

点导航 (Point Navigation)	在点导航中，智能体的任务是导航到离特定点一定固定距离内的任何位置。	
	一般来说，智能体在环境中的原点 (0, 0, 0) 初始化，固定目标点由相对于原点/初始位置的三维坐标 (x, y, z) 指定。为了顺利完成任务，人工智能体需要具备多种技能，如视觉感知 (visual perception)，事件记忆构建 (episodic memory construction)，推理/规划 (reasoning/planning) 和导航 (navigation)。人工智能体通常配有 GPS 和指南针，可以获取其位置坐标，并包含其相对于目标位置的方位。目标的相对目标坐标可以是静态的 (即只在事件开始时给出一次)，也可以是动态的 (即在每个时间步给出)。	
	采用端到端方法，在现实的自主导航环境中，利用不同的感官输入解决点导航问题。	Benchmarking Classic and Learned Navigation in Complex 3D Environments, 2019
	基础导航算法是 Direct Future Prediction (DFP)，其中相关输入 (如彩色图像、深度图和最近四次观测的行动) 由适当的神经网络 (如用于感官输入的卷积网络) 处理，并连接到双流全连接行动预期网络中，输出是对所有行动和未来时间步预测的未来测量预测。	Learning to Act by Predicting the Future, 2016
	Belief DFP (BDFP) 旨在通过在未测量预测中引入类似于中间地图的表征，使 DFP 的黑箱策略更具可解释性。其灵感来自神经网络中的注意力机制，以及强化学习中的后理表征和特征。	Benchmarking Classic and Learned Navigation in Complex 3D Environments, 2019
	SplitNet 提供了一种更模块化的方法。对于点导航，SplitNet 的架构由一个视觉编码器器和多个解码器组成，用于不同的辅助任务 (如自我运动预测) 和策略。这些解码器旨在学习有意义的表征。采用相同的 PPO 算法和用于策略训练，SplitNet 可在以前未见过的环境中轻松向策略转移方法。	SplitNet: Sim2Sim and Task2Task Transfer for Embodied Visual Navigation, 2019
	在室内环境中同时进行地图绘制和目标驱动导航。作者在 MapNet 的基础上增加了具有语义特征的 2.5D 记忆，并为导航策略训练了一个 LSTM。在以前未见过的环境中，这种方法优于不带地图的 LSTM 策略。	Simultaneous Mapping and Target Driven Navigation, 2019

分散分布式近端策略优化，具有广义优势估计功能，适用于资源密集型模拟环境中的分布式强化学习。		DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames, 2019
通过辅助任务提高采样和时间效率，改进 DD-PPO 的资源密集问题		Auxiliary Tasks Speed Up Learning PointGoal Navigation, 2020
Habitat Challenge 2019 冠军解决方案：结合了经典方法和基于学习的方法，因为基于端到端学习的方法计算成本高昂。这项工作以模块化方式将学习输入 "classic navigation pipeline"，从而将策略和感知知识融合地纳入低级导航。该架构由 learned Neural SLAM 模块，全局策略，局部策略和 analytical path planner 组成。Neural SLAM 模块利用观测和传感器预测地图和智能体状态估计值。全局策略始终将目标坐标输出为长期目标，并通过分析路径规划器将其转换为短期目标。最后，对局部策略进行训练，使其导航至该短期目标。模块化设计和分析规划器的使用有助于大大减少训练过程中的探索开销。		Learning to Explore using Active Neural SLAM, 2020

在物体导航中，智能体的任务是导航到一个特定类别的物体。

物体导航 (Object Navigation)	物体导航的基本思想是在一个未开发的环境中导航到一个由基标指定的物体。	
	智能体将在一个预定义位置初始化，任务是在该环境中找到一个物体类别的实例。物体导航通常比点导航更为复杂，因为它不仅需要许多相同的技能，如视觉感知和事件记忆 (episodic memory) 构建，还需要语义理解。这些使得物体导航任务更具挑战性，但同时也更有价值。	
	物体导航任务可以通过自适应 (adapting) 来演示学习。这有助于在没有任何直接监督的环境中实现导航的泛化。SAVN 通过元强化学习方法借鉴了这一点，因为智能体学习了一种目前监督的交互损失，这有助于驱动有效的导航。	Learning to Learn How to Learn: Self-Adaptive Visual Navigation Using Meta-Learning, 2018
	另一种方法是在执行导航规划之前学习物体之间的关系。这项工作实现了一个物体关系图 (ORC)，它不是来自外部的先验知识，而是在视觉感知所建立的假设图。该图由物体关系组成，如类别接近度 (category closeness) 和空间相关性 (spatial correlations)。	Learning Object Relation Graph and Tentative Policy for Visual Navigation, 2020
	有先验的导航侧重于以多模态输入 (如知识图谱或高保真输入) 的模式注入语义知识或先验，或在可见和不可见的区域中辅助训练人工智能智能体的导航任务。过去的研究表明，人工智能智能体可以利用与人类类似的语义功能先验知识，帮助智能体学习导航，并在看不见的区域中找到看不见的物体。	Visual Semantic Navigation using Scene Priors, 2018

人类能够感知并辅助指导智能体与物体物理位置之间的对应关系，从而执行导航以到达信号源。在这项工作中，人工智能体采集多种感官观测数据，如目标物体的视觉和声音信号，并找出从其起始位置到声音来源的最短导航路径。这项工作通过视觉感知和听觉，声音感知模块和动态路径规划器来实现。		Look, Listen, and Act: Towards Audio-Visual Embodied Navigation, 2019
一种让智能体学习按照自然语言指令在环境中导航的任务。这项任务的挑战在于理解感知视觉场景和语言。VLN 仍然是一项具有挑战性的任务，因为它要求智能体根据过去的行动和指令预测未来的行动。此外，机器人可能无法将自然语言指令与指令对齐。虽然视觉指导导航和视觉问答 (VQA) 密切相关，但这两项任务有很大不同。这两项任务都可以表述为基于视觉的序列到序列映射问题。不过，VLN 序列更长得多，需要不断输入视觉数据，并能提供摄像机视图，而 VQA 只需输入一个问题并生成答案。		
视觉对话导航 (Vision-and-Language Navigation, VLN)	辅助推理导航 (Auxiliary Reasoning Navigation) 解决了四个辅助推理任务：轨迹预测 (trajectory retelling)，进度估计 (progress estimation)，角度预测 (angle prediction) 和跨模态匹配 (cross-modal matching)。智能体学会对之前的行动进行推理，并预测任务的未来信息。	Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks, 2019
	视觉-对话导航 (Vision-dialog navigation) 目的是训练机器人发展与人类进行持续自然语言对话的能力，以辅助其导航。目前在这一领域开展的工作使用了跨模态记忆网络 (CMN)，通过独立的语言记忆和视觉记忆模块来记忆和理解与过去导航行动相关的有用信息，并进一步利用这些信息做出导航决策。	Vision-Dialog Navigation by Exploring Cross-Modal Memory, 2020

(2)评估指标

- 1)(反归一化)路径长度加权成功率(success weighted by (normalized inverse) path length, SPL)
- 2)成功率
- 3)路径长度比(pathlength ratio)
- 4)成功距离/导航误差(distancetosuccess/navigation error)

评估 VLN 智能体：

- 1)oracle success rate.即智能体沿轨迹停在离目标最近点的比率;
- 2)轨迹长度(trajectorylength)

视觉对话导航：

- 1)目标进度(goal progress):即智能体向目标位置前进的平均进度。
- 2)oracle path success rate,即智能体沿最短路径停在离目标最近点的成功率。
- 3)成功率。
- 4)oracle success rate。

(3)数据集

Matterport3D和Gibson V1

VLN 需要一种不同的数据集：Matterport3D模拟器的Room-to-Room(R2R)数据集

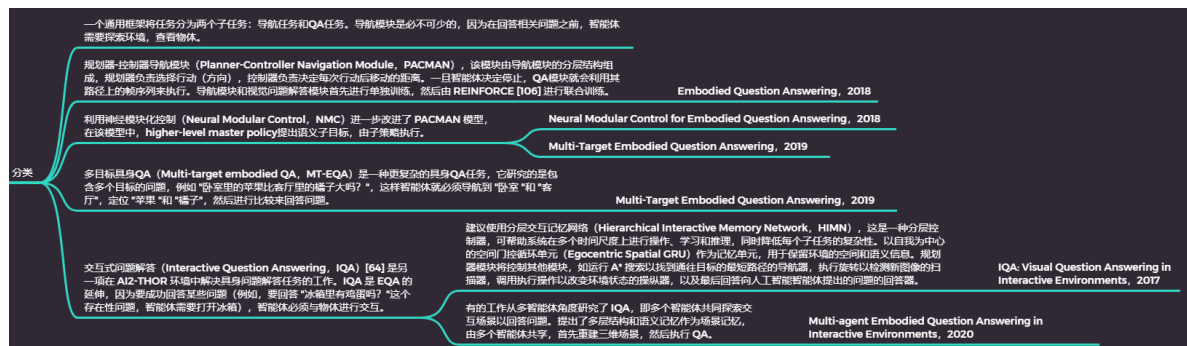
视觉对话导航：Cooperative Vision-and-Dialog Navigation(CVDN)数据集



## C. Embodied Question Answering

在物理具身状态下执行 QA，人工智能智能体需要具备广泛的人工智能能力，如视觉识别、语言理解、问题解答、常识推理、任务规划和目标驱动导航等。因此，具身问答可以说是目前具身人工智能研究中最繁重、最复杂的任务。

### (1)方法



### (2)评估指标

评估指标：具身QA和IQA涉及两个子任务：1) 导航，2) 问题解答，这两个子任务的评估基于不同的指标。

导航性能通过以下方面进行评估：

- 1) 导航终止时与目标的距离，即导航误差 (dT)；
- 2) 从初始位置到最终位置与目标距离的变化，即目标进度 ( $d\Delta$ )；
- 3) 在整个事件 (episode) 中任何一点与目标的最小距离 ( $dmin$ )；
- 4) 在达到最大事件长度之前，智能体为回答问题而终止导航的事件百分比 (%stop)；
- 5) 智能体在包含目标物体的房间内终止导航的问题百分比 (%rT)；
- 6) 智能体至少有一次进入包含目标物体的房间的问题百分比 (%re)；
- 7) 目标物体交并比 (IoU)；
- 8) 基于 IoU 的命中精度 (hT)；
- 9) 事件长度 (episode length)，即轨迹长度 (trajectory length)。

指标(1),(2)和(9)也用作视觉导航任务的评估指标。

QA性能通过以下方面进行评估：

- 1) 真实答案在预测中的平均排名 (mean rank, MR)；
- 2) 准确率。

### (3)数据集

EQA[61]数据集基于House3D，House3D是流行的SUNCG[33]数据集的一个子集，具有合成的房间和布局，类似于Replica数据集[107]。House3D 将 SUNCG 的静态环境转换为虚拟环境，在虚拟环境中，智能体可以在物理限制条件（如不能穿过墙壁或物体）下进行导航。为了测试智能体在语言基础、常识推理和导航方面的能力，[61] 使用 CLEVR [108] 中的一系列功能程序来合成有关物体及其属性（如颜色、存在性、位置和相对介词）的问题和答案。总共有 750 个环境中的 5,000 个问题，涉及 7 种独特房间类型中的 45 个独特物体。

对于 MT-EQA[63]，作者介绍了 MT-EQA 数据集，该数据集包含 6 种类型的组合问题，可比较多个目标（物体/房间）之间的物体属性（颜色、大小、距离）。

对于 IQA [64]，作者对一个大型数据集 IQUAD V1 进行了注释，该数据集由 75,000 道选择题组成。与 EQA 数据集类似，IQUAD V1 也有关于物体存在、计数和空间关系的问题。

# INSIGHTS AND CHALLENGES

---

- 见解:

- 通过图6中的相互连接,我们发现模拟器对于研究任务的适用性。Habitat-Sim和iGibson支持视觉探索和各种视觉导航任务,表明高保真度的重要性。然而,一些模拟器由于其独特特性,使它们更适合非体验智能独立任务,如深度强化学习,目前并不连接到任何体验研究任务。尽管如此,它们仍符合被分类为体验智能模拟器的标准。
- 对于体验问题回答和带先验的视觉导航等研究任务,需要体验智能模拟器具有多状态对象属性,这是由于这些任务的互动性质所决定的。因此,AI2-THOR无疑是首选的模拟器。最后,VLN是目前唯一不使用任何九个体验智能模拟器之一,而是使用Matterport3D模拟器的研究任务。这是因为以前的VLN工作不需要模拟器中的互动特性;因此Matterport3D模拟器足够。然而,随着VLN任务的进一步发展,我们可以预期VLN任务中需要互动性,因此需要使用体验智能模拟器。

- 挑战:

- 在体验智能模拟器方面存在着现实主义、可扩展性和互动性等方面的挑战。现有的体验智能模拟器在功能性和保真度方面已经达到了一定水平,使它们与传统用于强化学习的模拟器有所区别。然而,尽管体验智能模拟器的差异性不断增加,但在现实主义、可扩展性和互动性等方面仍存在一些挑战。
- 现实主义方面,模拟器的视觉保真度和物理特性是机器人社区追求的理想测试平台,如导航和交互任务。然而,缺乏同时具备基于世界场景和先进物理特性的体验智能模拟器。对于物理特性,需要具有先进物理特性的体验智能模拟器,以提供训练体验智能代理执行具有复杂物理交互的任务的理想测试平台。然而,目前只有一个符合这一标准的模拟器,即ThreeDWorld。因此,缺乏具有布料、流体和软体物理等先进物理特性的体验智能模拟器。我们相信,3D重建技术和物理引擎的进步将提高体验智能的现实主义水平。
- 可扩展性方面,与从众包或互联网轻松获取的基于图像的数据集不同,收集大规模基于世界场景的3D场景数据集和3D对象资产的方法和工具相对稀缺。目前,收集现实感3D场景数据集的方法需要通过摄影测量学扫描物理房间,例如Matterport 3D扫描仪、Meshroom或甚至移动3D扫描应用程序。然而,这些方法对于收集大规模3D对象和场景扫描并不商业可行,主要是因为用于摄影测量的3D扫描仪昂贵且不易获得。因此,可扩展性的瓶颈在于开发用于大规模收集高保真度3D对象或场景扫描的工具。希望通过3D学习方法的进一步发展,我们将能够扩大大规模3D数据集的收集过程。
- 互动性方面,体验智能模拟器中与功能性对象进行精细操作的能力对于复制人类与真实世界对象的交互至关重要。大多数基于游戏场景的模拟器提供精细的对象操作能力和符号交互能力,但由于游戏场景模拟器的性质,许多在此环境中进行的研究任务将选择其符号交互能力,而不是精细的对象操作。另一方面,基于世界场景的模拟器的代理具有粗大的运动控制能力,而不是符号交互能力。然而,这些模拟器中的对象属性主要是在表面上可交互的,这允许进行粗大的运动控制,但缺乏多状态对象类,即对象具有的状态变化数量。因此,需要在对象功能性和对象属性以及体验智能代理在环境中执行的动作复杂性之间取得平衡。