

# 具身导航：从模拟到现实

报告人：蒋树强

2024-09-29



中国科学院计算技术研究所  
Institute of Computer Technology, Chinese Academy of Sciences

# 汇报大纲

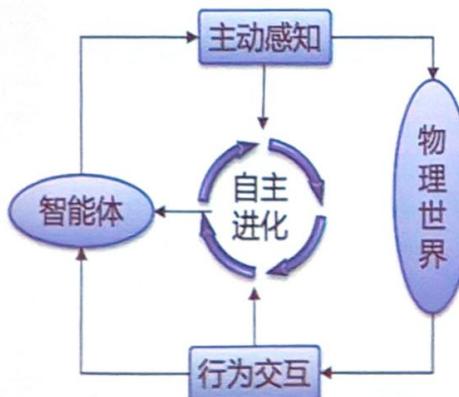


# 具身智能

## 具身智能：通过身体与环境互动来实现的智能

智能体在物理世界感知和理解环境，通过适应性行为和自主学习来完成任务

### 具身智能是人工智能的初心起点



阿兰图灵：  
《Computing Machinery and Intelligence》 1950

### 具身智能是人工智能的国际前沿

- 2023美国《国家人工智能研发战略计划》中包括3项有关具身智能
- 欧盟“地平线2020”框架计划“GOAL-ROBOTS”项目
- 国际大公司及知名大学在具身智能方向都有具体布局



### 具身智能是人工智能的万众焦点

- 2023年3月Google发布多模态具身视觉语言模型PaLM-E
- 2023年10月，DeepMind发起34个机构共同启动RT-X
- 2024年1月，New York University发布OK-Robot
- 2024年3月，NVIDIA发布人形机器人基础模型Project GR00T、新款人形机器人计算机Jetson Thor



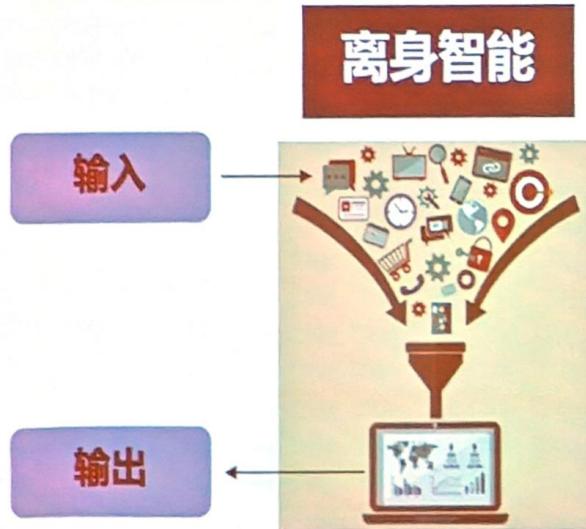
PaLM-E

Ok-Robot

人工智能从互联网走向物理世界，具身智能是核心关键

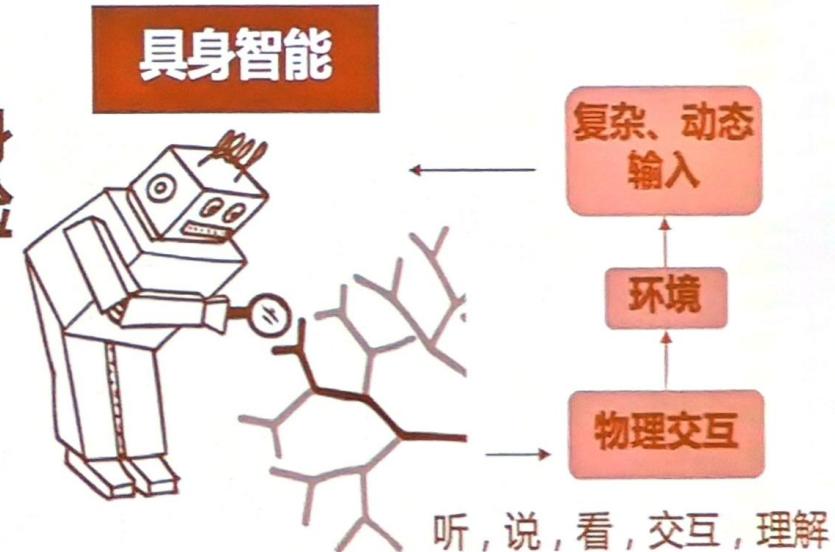
# 具身智能

具身智能：通过身体与环境互动来获得的智能



离身智能 ? 具身体验

VS.



单一的符号智能往往与真实世界相脱节，  
认知与身体解耦

被动

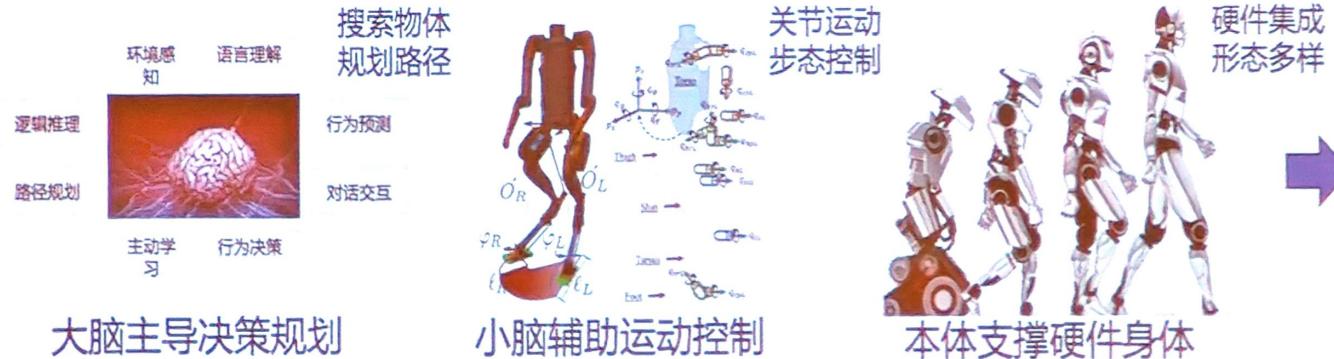
智能是具身化和情景化的，具身智能可通过  
与真实世界的交互完成任务

主动

智能与具身体验存在内在联系，需要将它们进行联合研究

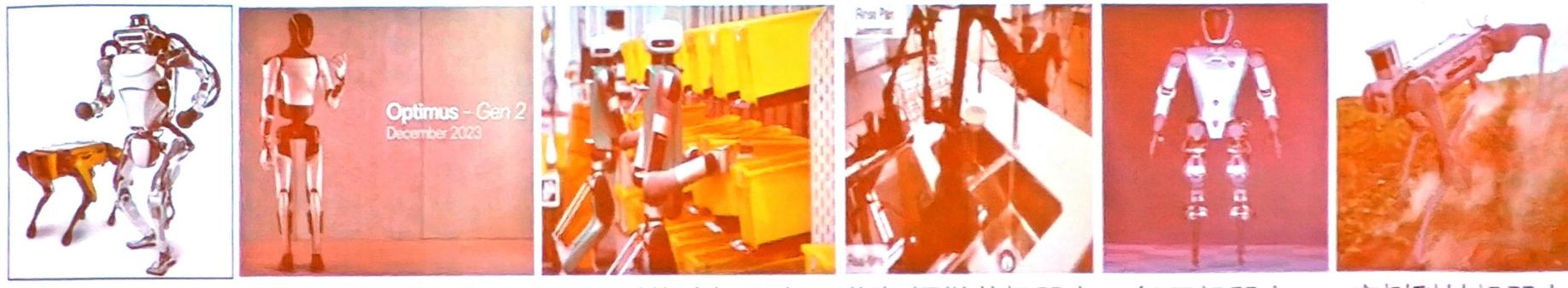
# 具身智能的本体支撑：机器人

**具身智能:大脑、小脑、本体紧密耦合、互相支撑**



智能化应用

**国内外机器人行业蓬勃发展，有力支撑具身智能广泛应用**



波士顿动力机器人 特斯拉人形机器人

亚马逊物流机器人

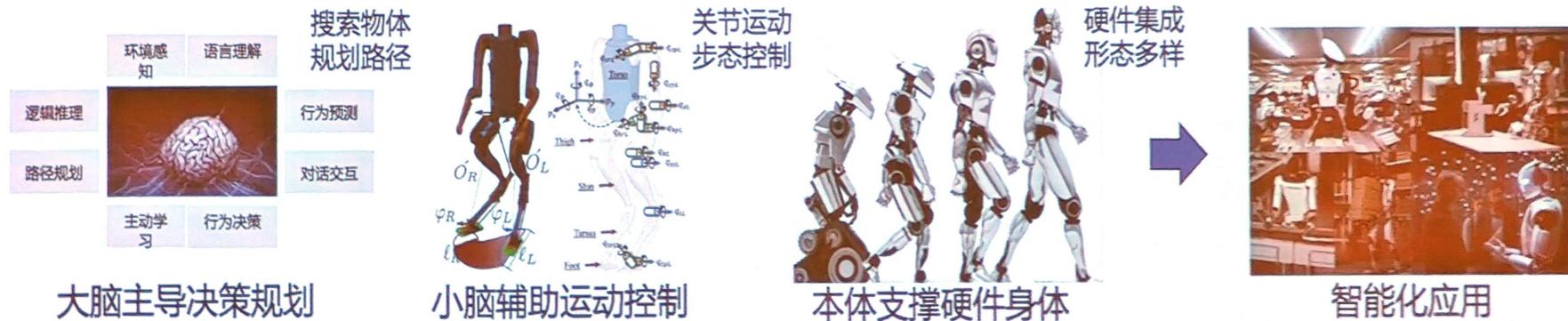
斯坦福做菜机器人

智元机器人

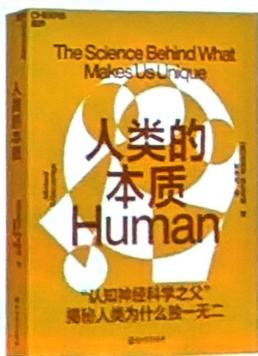
宇树科技机器人

# 具身智能的本体支撑：机器人

**具身智能: 大脑、小脑、本体紧密耦合、互相支撑**



- 大脑这一器官，把我们跟其他所有物种区别开来，我们在肌肉和骨骼方面并没有什么特别的，但我们的大脑与众不同。



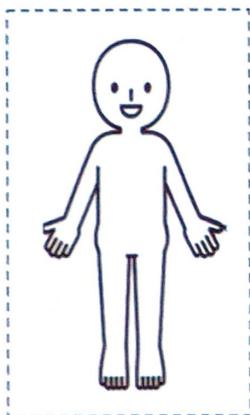
帕什科·拉基奇 ( Pasko T. Rakic )

美国耶鲁大学医学院神经科学家

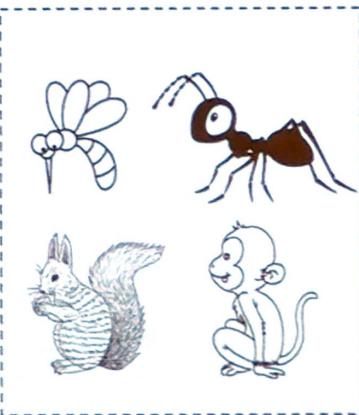
# 智能体

- 智能体：可以被看做通过感应器感知其环境并通过效应器来作用环境的任何事物。

■ 形式：



人



动物



机器人

区别于



非智能体

■ 基本特征：**自主性**   **感知**   **推理与决策**   **行动**   **学习**

# 智能体的性质

- 服从于物理法则
- 通过运动（真实世界的交互）来产生感觉刺激
- 通过自己的行为影响环境
  - 通过身体、脑（控制系统）及环境达到稳定的吸引子状态
- 可进行身体的形态学计算（不需要大脑，身体适应性）



R.Pfeifer, J.bongard,  
《身体的智能》



# “生物体-人” 是综合的智能体

## ■ 人的智能

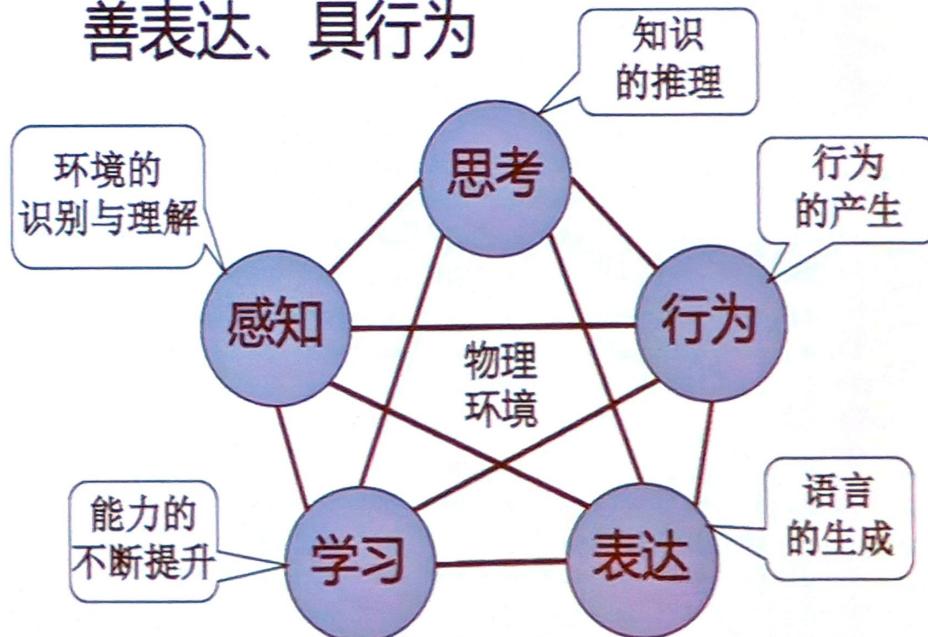
□ 知觉、学习、记忆、推理、语言理解、知识获得、情感、意识和动作控制等

## ■ 人是综合的智能体

智能是由系统的多部件之间的交互作用以及与环境交互作用所突现出来的  
总体行为

R Brooks 1999, Cambrian Intelligence: The Early History of the New AI, MIT Press, Cambridge, MA.

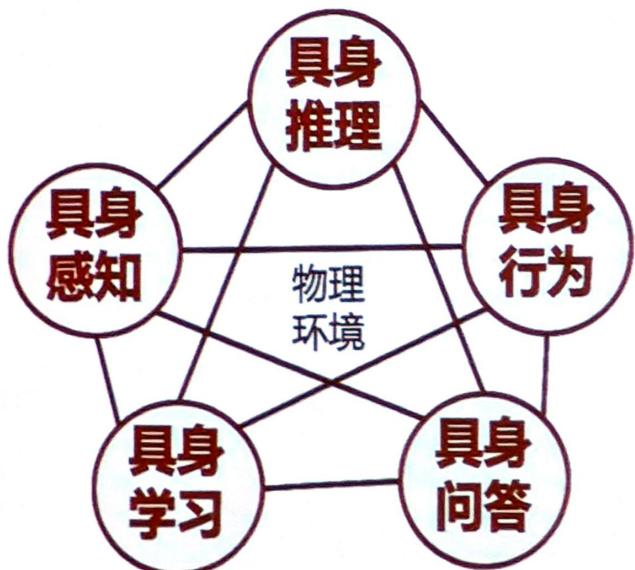
- 能感知、会思考、可学习、善表达、具行为



- 每个部分都相互关联
- 从各自为战到兼而顾之

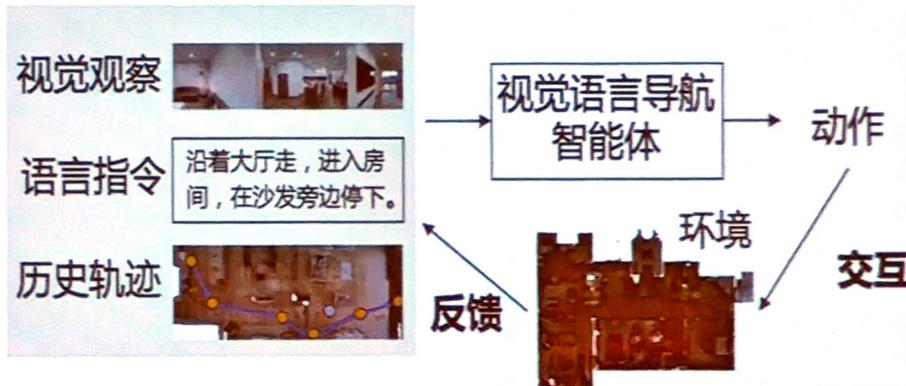
# 具身智能是综合的智能体

## 多任务互相促进



各项具身任务深度融合，紧密关联

## 举例：视觉语言导航



区域知识增强内容表示，促进多模态匹配 [CVPR23,ict]  
网格记忆存储时空关系，捕捉细粒度信息 [ICCV23,ict]

# 具身智能研究趋势

## 特定指令控制

- 语音指令



- 手势指令 [RSS 2023]



被动接收具体指令，完成  
特定设计行为

## 指令行为

指定具体指令

## 人机语言交互

- 细粒度的逐步语言指导  
[CVPR 2019 Best Paper]



- 交互式的人机问答协同  
工作 [ICRA 2023]

根据语言引导逐步完成复  
杂任务行为

## 语言引导行为

- 归纳还原
- Affordance探索等



主动交互  
建立对环境的认知

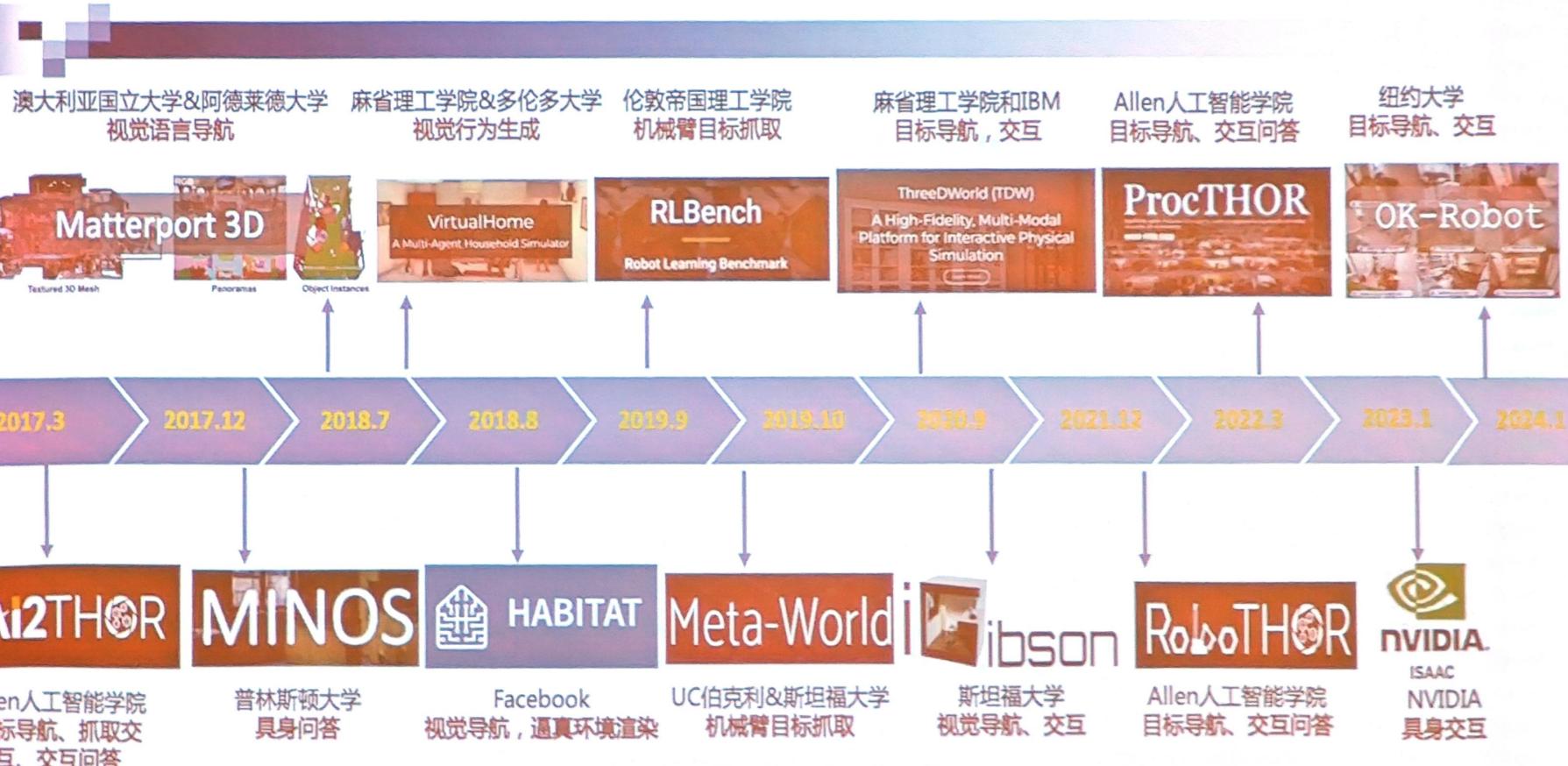
## 自主行为交互

独立探索作业

具身智能正在促进行为交互模型从被动接收指令向  
主动探索发展



# 研究现状



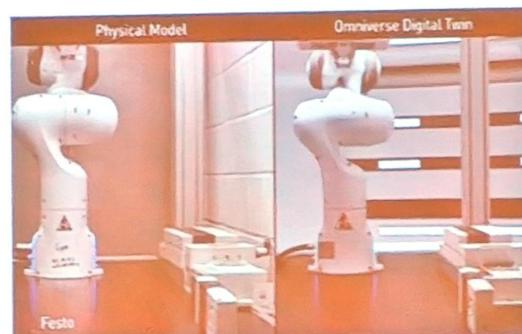
视觉导航是具身智能的重要组成部分

# 虚拟环境与现实环境

- 现实环境训练成本大
- 虚拟环境中可以进行大规模的训练（深度学习，强化学习）
- 交互性强、便于评测
- 从虚拟到现实环境迁移



货物导航机器人



抓取机器人



自动驾驶

虚拟环境和具身智能紧密交织在一起

真实环境

虚拟环境

真实环境

# 研究现状



# 具身大模型



# 具身多模态大模型

**当前现状**

(1) 以语言大模型为中心的具身系统，如VoxPoser、EmbodiedGPT、TidyBot

```

    graph LR
        视觉观察[视觉观察] --> 视觉语言模型[视觉语言模型]
        视觉语言模型 -- 规范化文本 --> 语言大模型[语言大模型]
        语言大模型 -- 行为规划 --> 执行器[执行器]
        执行器 --> 行为动作序列[行为动作序列]
        语言大模型 -- 回答文本序列 --> 回答文本序列
    
```

(2) 视觉文本数据、机器人演示数据联合训练的具身系统，如PaLM-E、RT-X

```

    graph LR
        视觉观察[视觉观察] --> 视觉语言模型[视觉-语言-动作模型]
        语言指令[语言指令] --> 视觉语言模型
        状态参数[状态参数] --> 视觉语言模型
        视觉语言模型 -- 回答文本序列 --> 回答文本序列
        视觉语言模型 -- 行为动作序列 --> 行为动作序列
        语言嵌入空间[语言嵌入空间] <--> 视觉语言模型
    
```

**未来趋势**

VoxPoser机械臂操作演示：  
Open the top drawer, and watch out for that vase.

感知难

- (1) 视觉、听觉、触觉的多感官协同

学习难

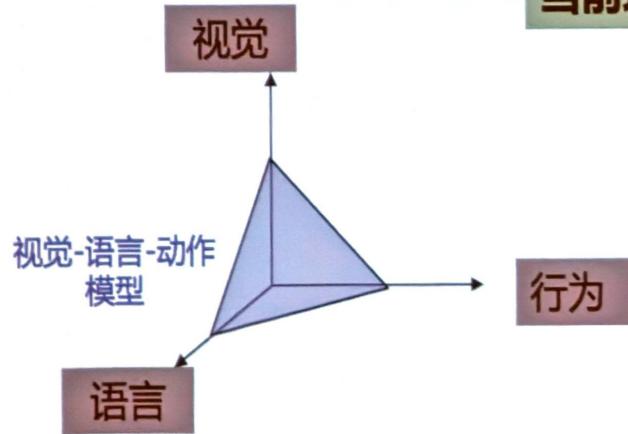
- (2) 结构化知识模式
- (3) 实时行为反馈调节与长期行为经验学习

决策难

- (4) 小样本人类示教学习
- (5) 心智理论与人类意图理解

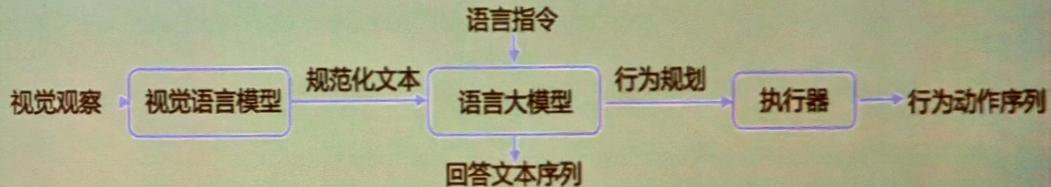
...

# 具身多模态大模型

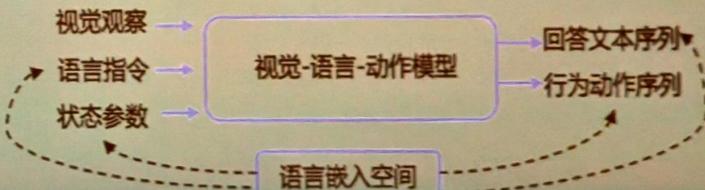


## 当前现状

(1) 以语言大模型为中心的具身系统，如VoxPoser、EmbodiedGPT、TidyBot

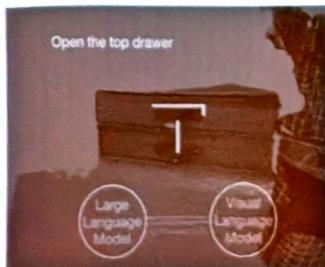


(2) 视觉文本数据、机器人演示数据联合训练的具身系统，如PaIM-E、RT-X



VoxPoser机械臂操作演示：

Open the top drawer, and watch out for that vase.



## 未来趋势

感知难 → (1) 视觉、听觉、触觉的多感官协同

感知难 → (2) 结构化知识模式

学习难 → (3) 实时行为反馈调节与长期行为经验学习

决策难 → (4) 小样本人类示教学习

决策难 → (5) 心智理论与人类意图理解

...

# 数据积累是具身大模型的重要支撑

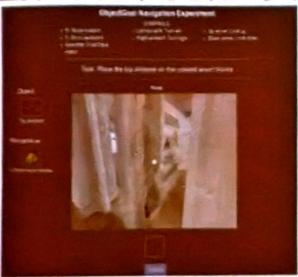
口数据采集与获取困难，数据量相对受限

## 虚拟式

- 建虚拟仿真环境，来模拟现实环境
- 在虚拟环境中训练智能体

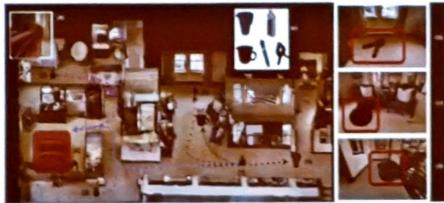


- 虚拟环境中收集人类演示



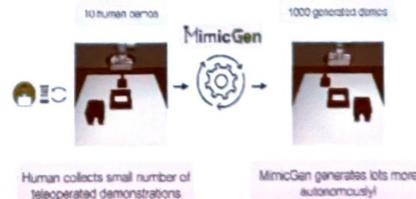
## 生成式

- 训练数据生成



- 根据人类演示数据，生成更多训练数据

MimicGen generates large datasets from few human demos



## 网络式

- 互联网数据预训练，学习通用知识
- 具身数据微调，学习动作控制

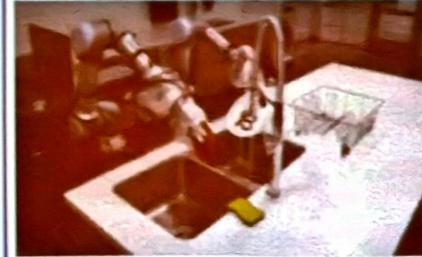


## 表演式

- 少量人类演示

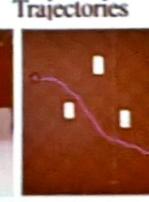
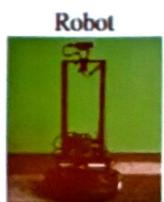
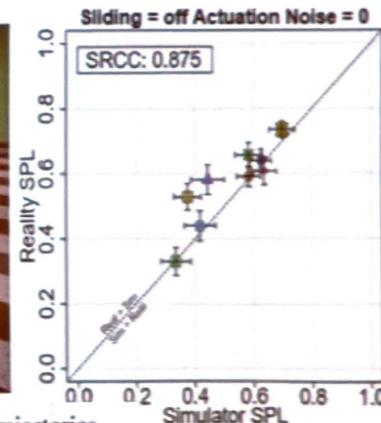


- 机器人从收集的数据中学习



# 虚拟到现实

- 具身智能体是在仿真环境中进行的，而不是在真实环境中进行
- 通过多模态传感器进行感知、交互和决策，形成综合的空间认知和操作能力
- 利用模拟器中的缺点和偏差(imperfections and biases) 实现强大的性能，这可能无法代表现实中的性能

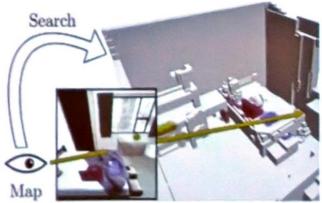
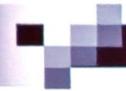


真实环境

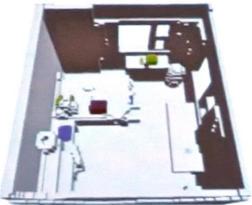
模拟器

真实环境和模  
拟器的相关性

# 真实世界 VS 虚拟世界



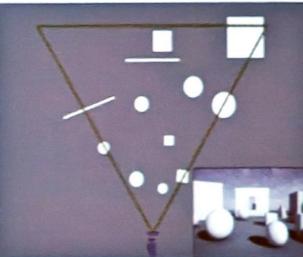
VS



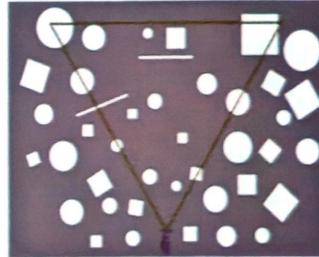
信息交互需通过智  
能体完成

真实世界的智能  
体是具身的

真实世界不能以明  
确的离散状态描述



VS



只能看到视觉范围  
内的东西

需要并行处理多个  
任务

智能单体需要同时  
做多件事情

真实世界  
的特点

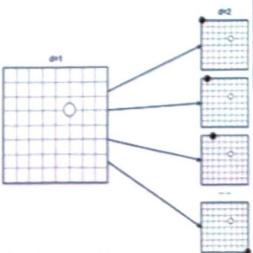
智能体能获得的  
信息很有限

实际装置容易受干  
扰或发生故障

信息具有噪声和不  
确定性



VS



真实世界有自身的  
动态机理



VS



# 具身智能：虚拟到现实

## 可控具身性仿真能力

单模态准确响应

多模态综合作用

挑战

高精度离散预测

连续动态变化

虚拟

现实

- **多模态传感器仿真**：共享底层模型和数据生成能力
- **动态过程仿真**：有效表示智能体位置及部件操作过程



## 多任务能力

视觉导航  
目标抓取  
归纳还原

单个智能体  
挑战  
多任务需求

虚拟

现实

- **世界模型**：构建多任务的通用模型
- **逆图形**：反投影真实世界到游戏引擎模拟器



## 资源适应能力

高性能服务器

准确仿真传感

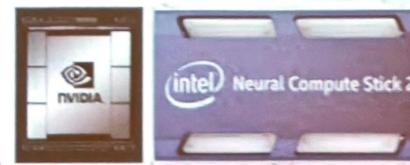
虚拟

边缘计算设备

有噪音真实传感

现实

- **专业边缘计算设备**：智能芯片、边缘加速设备
- **更真实的仿真环境**：引入真实噪音、生成真实数据



## 低成本测试能力

环境数据可定制

通用代码接口

虚拟

动态环境难控制

个性化接口定制

现实

- **环境数据迁移**：域适应，对抗生成机制
- **通用开源机器人**：LocoBot, TurtleBot



# 具备智能任务所需计算资源



任务	模拟器训练			真实环境应用
	目标导航	归纳还原	视觉语言导航	
数据集\平台	MP3D\Habitat	ProcTHOR\AI2THOR	R2R,REVERIE\Matterport3D	LocoBot, TurtleBot, Aloha
计算资源	4 RTX 3090 24GB	16 Tesla A100 40GB	20 Tesla A100 40 GB	末端CPU，推理GPU
预处理	60h预训练	96h预训练	20h预训练	5min建图

MP3D



40类物体  
训练：61公寓；  
验证：11公寓；  
测试：18公寓；

ProcTHOR



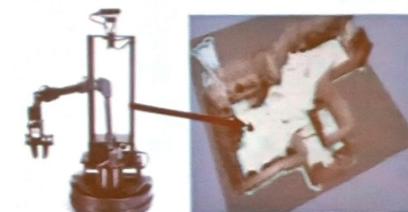
10,000公寓,打乱1~5个物体  
预训练：2500公寓  
验证：20公寓  
微调：80公寓 (AI2THOR)  
测试：20公寓 (AI2THOR)

R2R



90个场景，21567条指令  
训练：61个场景，14025条指令  
验证：见过场景61，未见过11  
测试：18个场景，4173条指令

真实环境



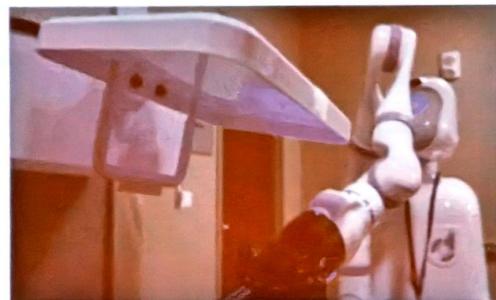
真实多房间场景，6种房间  
实时栅格地图构建\更新  
测试：目标导航

# 具身机器人



具身机器人未来有广阔应用前景：帮助人、代替人、增强人

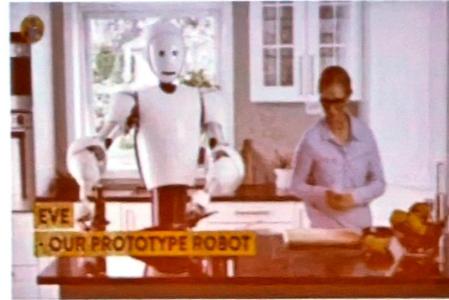
- 家庭服务
- 养老服务
- 办公服务
- 餐饮服务
- 军事应用
- 工业应用
- 教育应用
- 元宇宙



家庭\养老服务



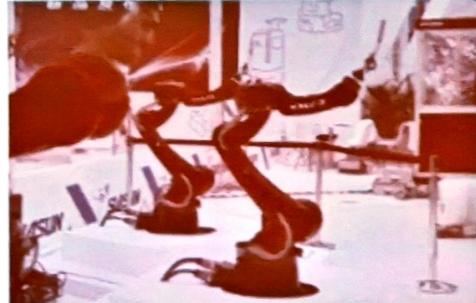
办公服务



餐饮服务



军事应用



工业应用



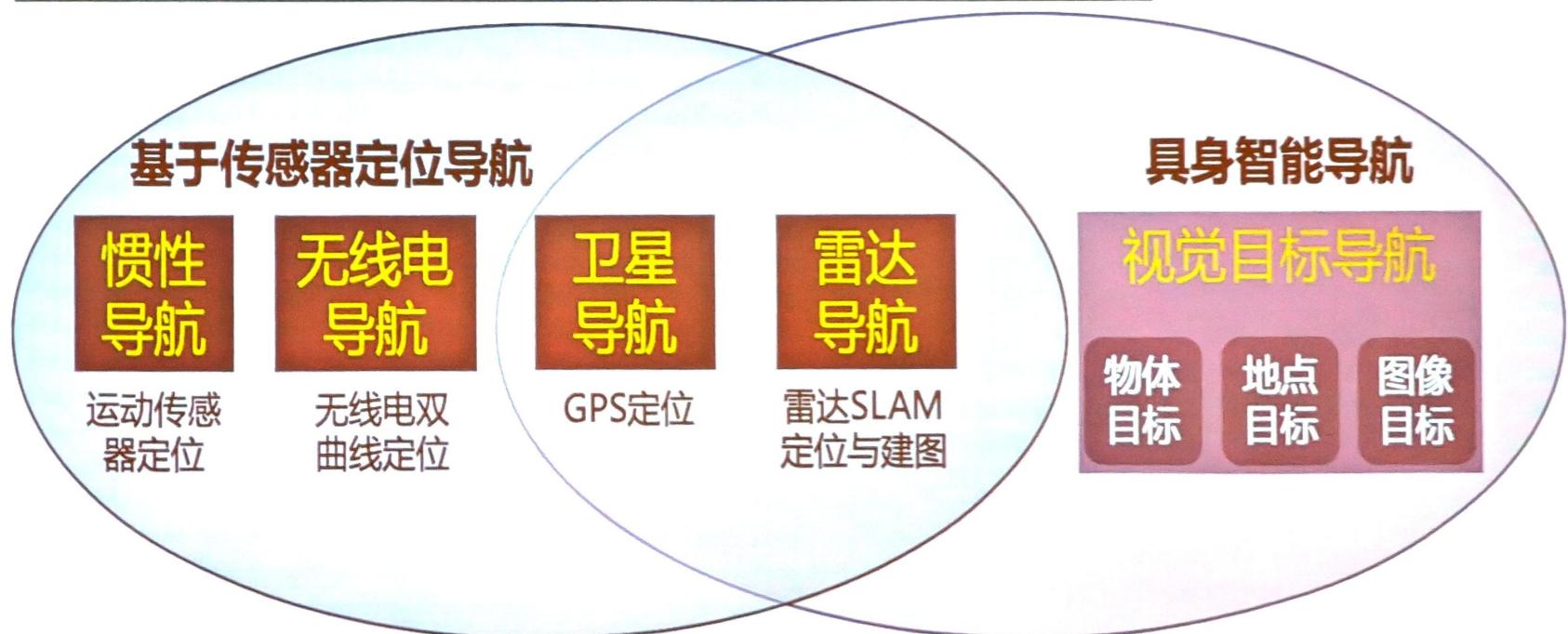
教育

# 汇报大纲



# 导航技术研究现状

导航：以某种方式引导人或航行体到达目标地点



传统导航以传感器定位为核心，应用门槛较高；具身智能导航以基于视觉的智能模型为基础，应用灵活性强，但研究挑战性高



# 具身导航：让机器人自己找到路

## 具身导航任务

给定目标或语言指示，智能体通过实时视觉感知环境，进行路径规划以找到目标

### 实时多模态信息输入

#### 视觉信息-实时观测

视觉传感器：颜色、纹理、形状等

#### 深度信息-实时观测

激光雷达/深度相机：场景三维结构等

#### 目标信息-任务定义

地点目标/物体目标/图像目标/语言描述

#### 语义信息-辅助知识

知识图谱：场景、物体、属性等



### 多样式目标定义

#### 地点式目标

“东北方向2m”

#### 物体式目标

“茶几”

#### 图像式目标



#### 语言描述式目标

左转，沿着大厅走，直到你  
右边找到一扇门。右转进入  
房间，在沙发旁边停下。

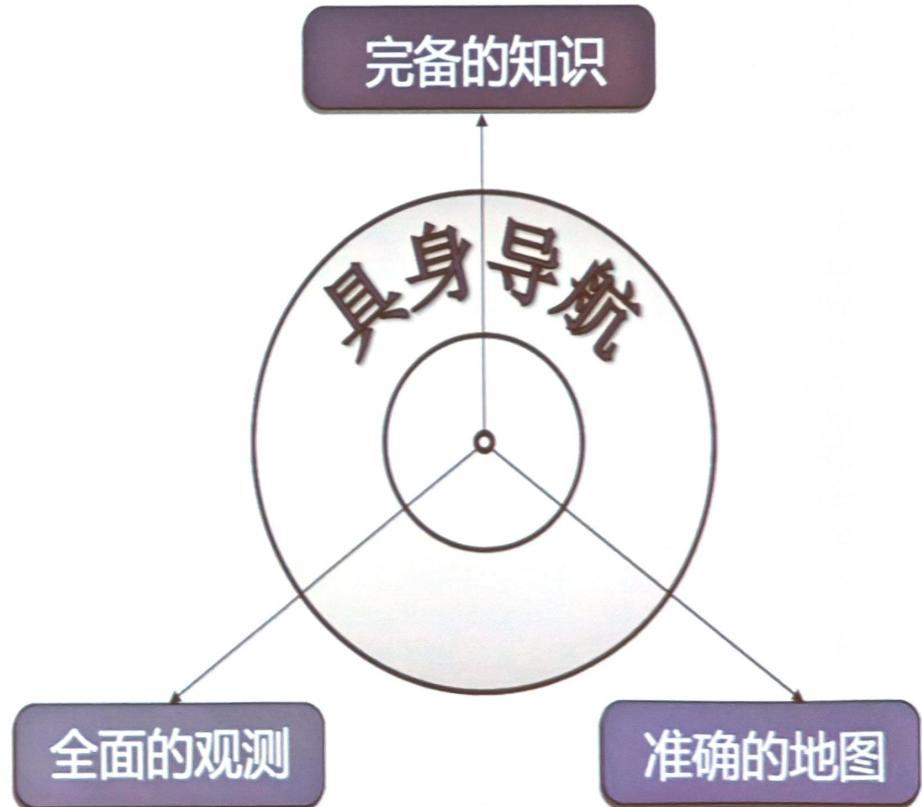
# 具身导航-让机器人在动态环境中自己找到路



# 具身导航是信息不完备情况下的开放环境导航

## ■ 几个例子

- 蚂蚁归巢
- 机器人送餐
- 人从复杂的迷宫里返回
- 人在家里找丢失的钥匙

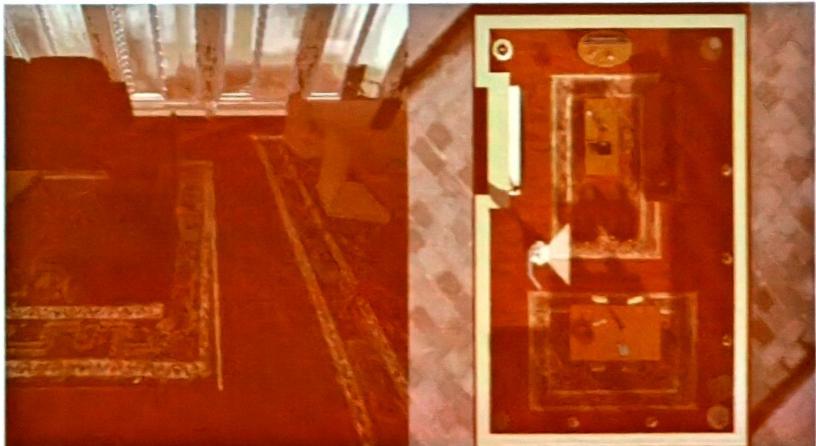


# 视觉目标导航

- 在未知的三维环境中
- 给定目标物体的语义（如：电视，沙发）
- 依赖第一人称RGB图像或多传感器信息
- 导航到指定目标物体

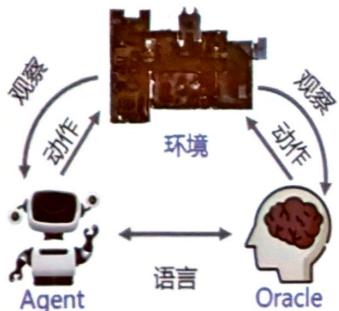


成功示例



导航示例

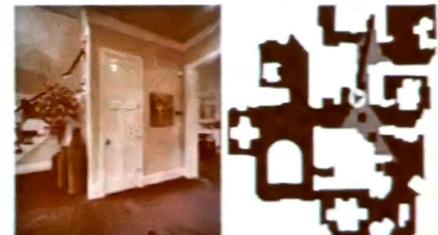
## 根据语言指令行走到指定位置



走向围栏，随后向右经过楼梯。走进起居室后右转，在桌子前停下。



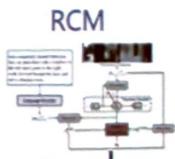
- (1) 理解语言
- (2) 关联视觉语言
- (3) 动作预测



Walk toward the railing and right past the stairs. Walk into the living room and turn right. Stop by the end table.



2018



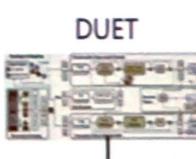
2019



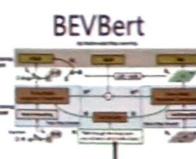
2020



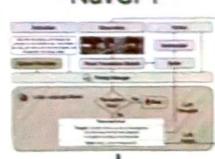
2021



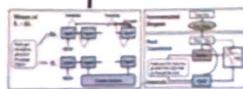
2022



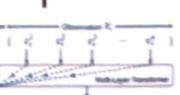
2023



NavGPT



EnvDrop



VLNBERT



GridMM

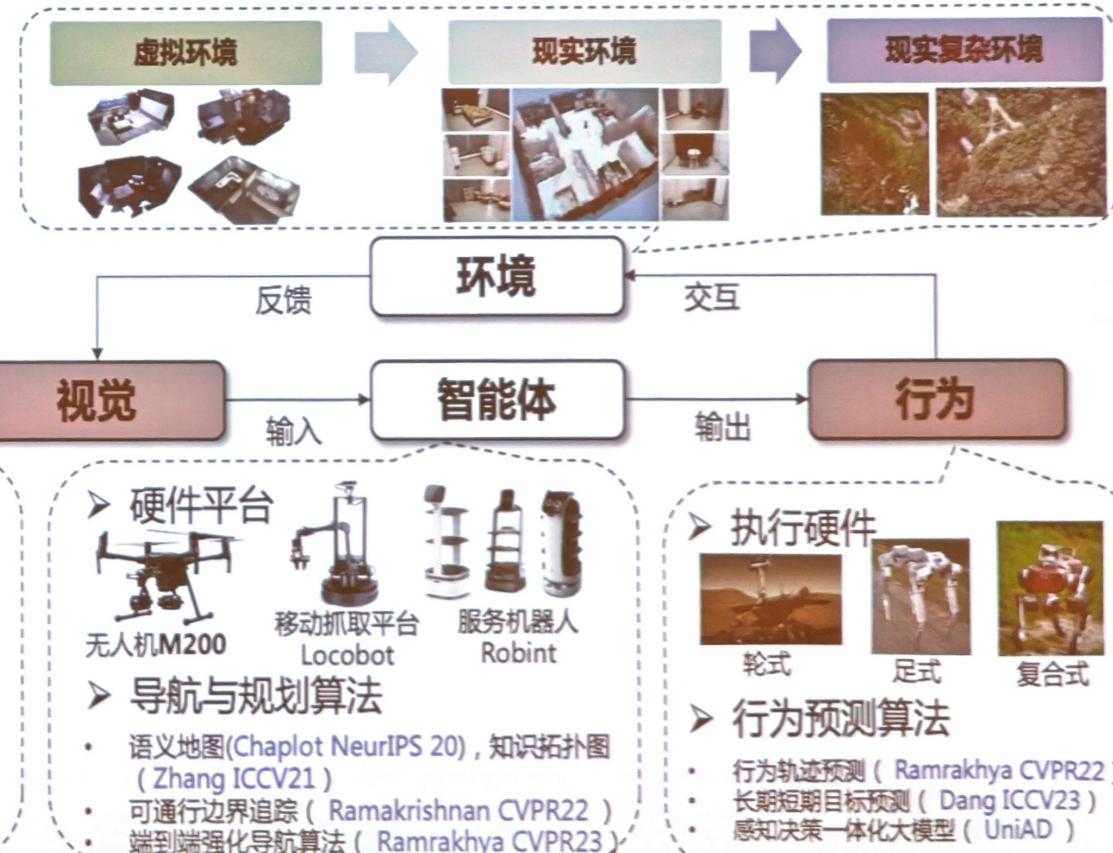
## 视觉导航：在未知场景下，通过视觉观测，导航到指定目标

### 视觉目标导航任务示例



RGB视图

第三视角俯视图



#### 感知硬件



RGB-D摄像头



激光雷达



臂端摄像头

#### 视觉感知算法

- 目标检测 (YOLO, DETR)
- 视觉分割 (SAM)
- 视觉预训练感知头 (CLIP, BEiT-3)



无人机M200



移动抓取平台 Locobot



服务机器人 Robint

#### 导航与规划算法

- 语义地图(Chaplot NeurIPS 20), 知识拓扑图(Zhang ICCV21)
- 可通行边界追踪(Ramakrishnan CVPR22)
- 端到端强化导航算法(Ramrakhya CVPR23)

#### 执行硬件



轮式



足式

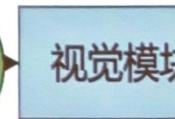


复合式

#### 行为预测算法

- 行为轨迹预测(Ramrakhya CVPR22)
- 长期短期目标预测(Dang ICCV23)
- 感知决策一体化大模型(UniAD)

# 视觉导航



- 视觉检测 DETR
- 视觉分割 SAM
- 多模态预训练头 CLIP

基于模块的目标导航技术

建图模块

- 度量地图[AUTON ROBOT 2002]
- 拓扑地图[ICLR 2018]
- 语义地图[ICRA 2021]

感知

预测航点模块

- 交互式学习[NIPS 2020]
- 监督式学习[CVPR 2022]

推断

动作



- 视觉感知结构[ICLR 2020, ICCV 2023]
- 注意力机制[CVPR 2021]

基于端到端的目标导航技术

关联学习

端到端的网络

策略函数

- 物体关联[ECCV 2020]
- 场景关联[ICCV 2019, 2021]
- 经验关联[CVPR 2019, 2023]
- 寻找/导航策略[ICCV 2023]
- 长短目标策略[IROS 2023]

感知

记忆

推断

动作

AI2THOR	RoboTHOR	Gibson	MP3D
120房间，单房间	90房间，复合房间	30别墅，多房间	67幢建筑，多房间
80%	50%	78%	40%

对环境的感知、记忆和路径规划是主要研究方向

# 视觉导航的挑战

## 未知场景下的视觉感知

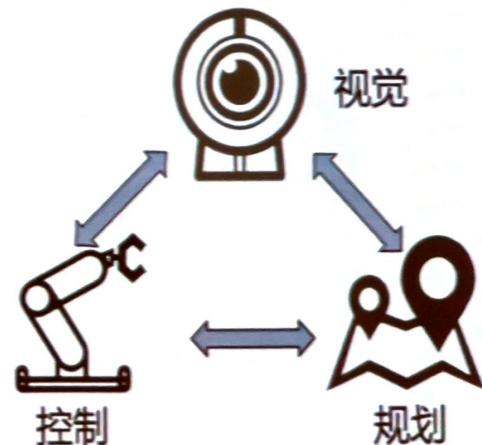
- 多种视觉信息的需求：
  - RGB，深度，检测，分割 .....
- 视觉数据差异：
  - 模拟器环境到现实环境
  - 未知环境 .....
- 开放标签的视觉感知：
  - 少样本的视觉目标
  - 零样本的视觉目标

## 未知场景下的路径规划

- 缺少全局地图：
  - 缺少全局SLAM地图
  - 如何学习布局的先验知识
  - 以何种形式存储知识
- 未知环境的泛化性：
  - 有限的训练房间和无限的未知环境
  - 知识的迁移和泛化
- 现实应用时传感器噪声：
  - 深度缺失，RGB模糊，检测分割误差

## 多智能单元的协同决策

- 多智能单元的协同决策
  - 视觉单元
  - 路径规划与决策单元
  - 执行单元



# 具身导航需要人类的高级认知能力



新任务：在新环境找罐“冰啤酒”？

## 初步观察试错



可视范围没有

## 认知理解

知识传递：冰啤酒一般储存于冰箱



空间推理：当前位置为客厅

## 规划并执行



① 导航到冰箱



② 开门



③ 拿“冰啤酒”

人类认知能力

空间推理

抽象思维

知识传递

学习适应

整合规划

心理模型

# 视觉导航/视觉语言导航需要综合的具身能力

## 多样指令：

实例级、未知物体、多物体导航

MM21, ECCV22, TIP23

## 理解环境：

场景知识图增强的具身导航

ICCV21, CVPR23

## 建立记忆：

基于网格记忆地图的视觉语言导航

ICCV23



## 利用知识：

知识增强的具身导航

CVPR23

## 预想将来：

探索和预想相结合的具身导航

CVPR24 (二维图生成), CVPR24 (三维图生成), NeurIPS24 (轨迹生成)

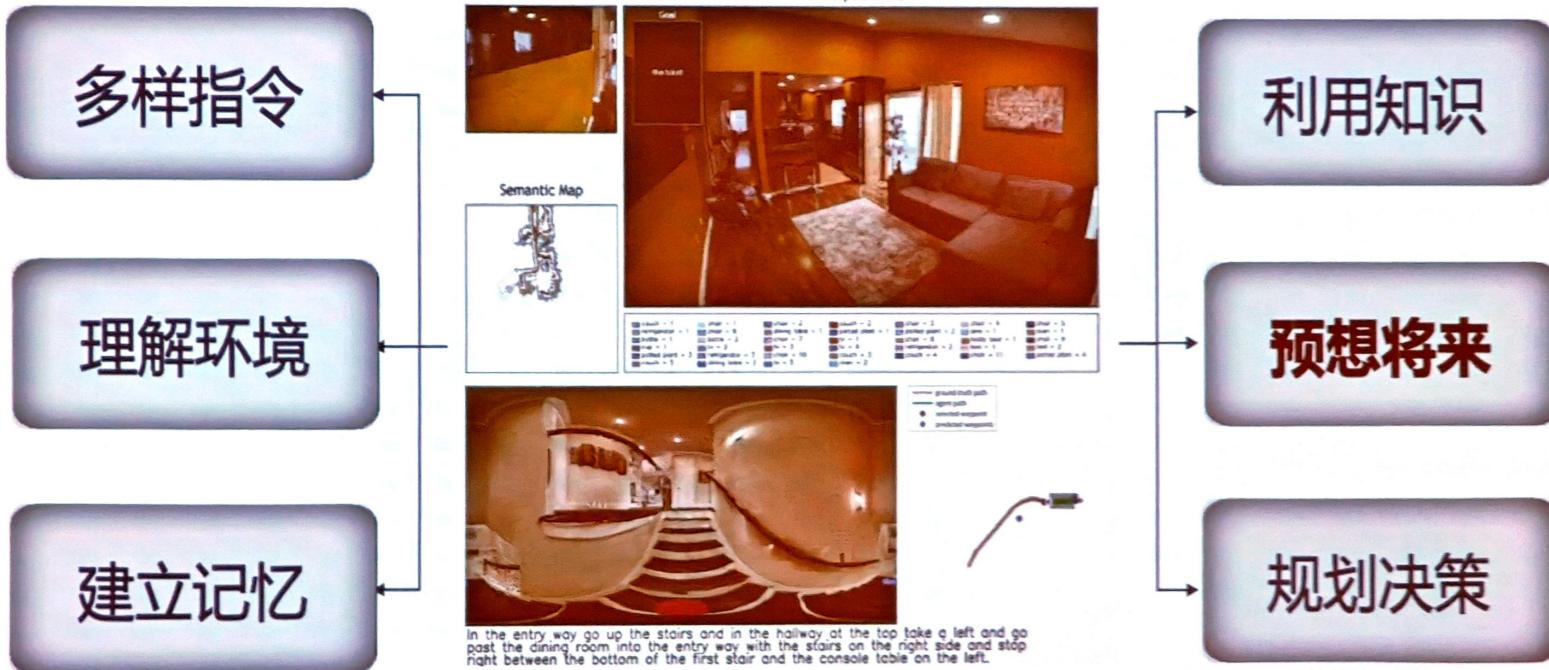
## 规划决策：

类别无关归纳还原/交互物体导航

NeurIPS23, CVPR24

- 建立心理地图是生物大脑在空间中导航的首选策略
- 人类导航中视觉和记忆信息占据主导，而很多导航能力卓越的动物都具有某些特殊的能力，如昆虫的偏振光视力、蝙蝠的空间记忆力和虹鳟鱼的磁感知能力等

# 视觉导航/视觉语言导航需要综合的具身能力



## 探索与预想相结合的具身导航

(二维图生成 : CVPR24、三维图生成 : CVPR24、轨迹生成 : NeurIPS24 )

# 探索与预想相结合的具身导航

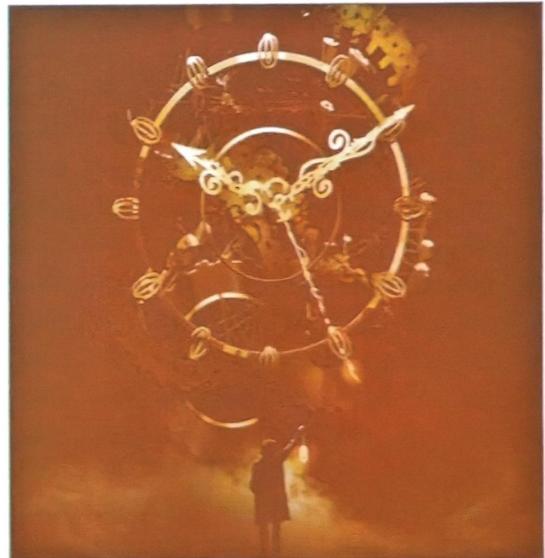
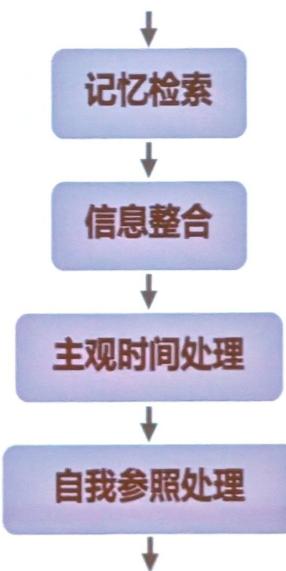
## 心理时间旅行

回忆历史事件，进行重建



预测未来事件，前瞻思维

《Mental time travel and the evolution of the human mind》



Suddendorf T, Corballis M C. Mental time travel and the evolution of the human mind[J].  
Genetic Social and General Psychology Monographs, 1997, 123(2): 133-168.

将自我投身到过去以预先经历未来

# 探索与预想相结合的具身导航



感知  
→



预测  
→



↑  
影响

未来事件

**推理：**层次结构，从感官输入到高阶抽象概念，信息双向流动。

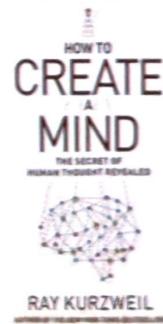
**记忆：**动态更新的数据库，存储过去模式、规则，而非事件累积。



《On Intelligence》

**PRTM理论（思维模式识别）：**

- 人类智能源于识别和预测模式的能力
- 大脑的设想会影响对事物的实际感知。

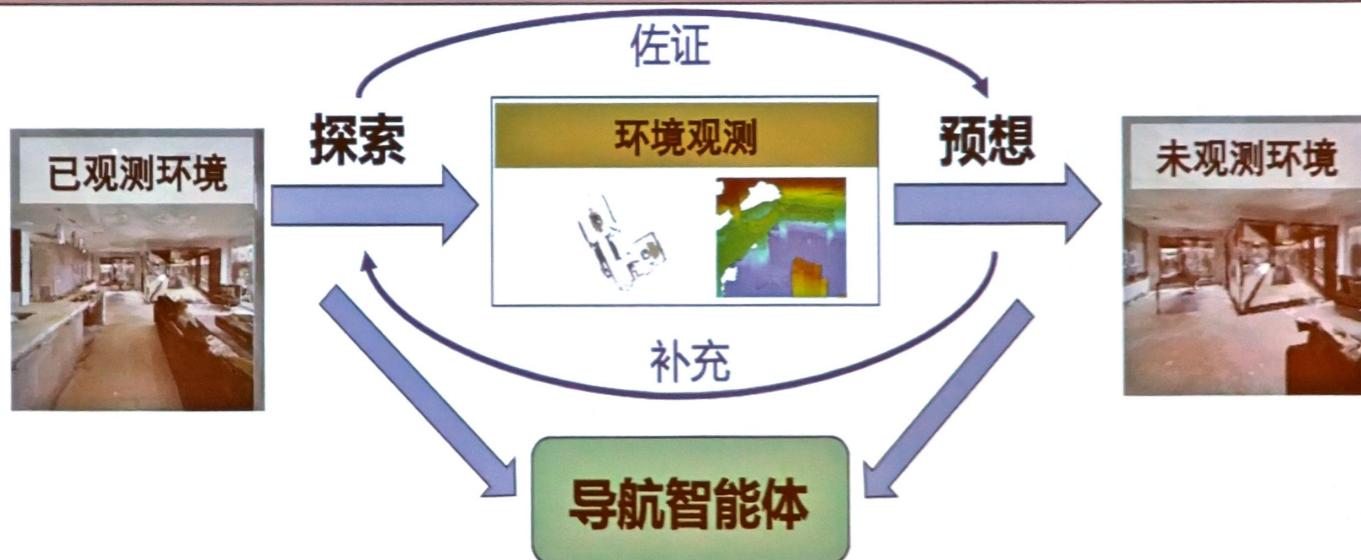


《How to Create a Mind》

智能的核心在于预测，同时预测影响对世界的感知

# 探索与预想结合的具身导航

探索采集实际环境信息，预想猜测未知环境信息，相互佐证与补充



## 基于自监督地图生成的目标导航

- 将探索过程中的历史观察编码为语义地图
- 利用地图补全模块预想未知区域地图，预测目标位置



## 基于神经辐射表征的视觉语言导航

- 将探索过程中的历史观察编码为三维表征
- 利用三维空间体积渲染预想未来场景，规划路径



# 研究内容-具身智能中的行为预测与导航

## 代表进展：基于自监督地图生成的物体视觉导航(CVPR2024)

### ➤ 挑战与动机

- 环境布局未知，智能体仅有局部观测信息

### ➤ 研究思路

- 基于自监督生成地图，从局部观测推断未探索区域信息

### ➤ 效果

- 公开模拟器 AI2Thor，RoboTHOR，Gibson 和 MP3D 达到业内最佳性能



# 研究内容-具身智能中的视觉语言导航

## 代表进展：基于神经辐射表征的前瞻探索策略(CVPR2024)

### ➤ 挑战与动机

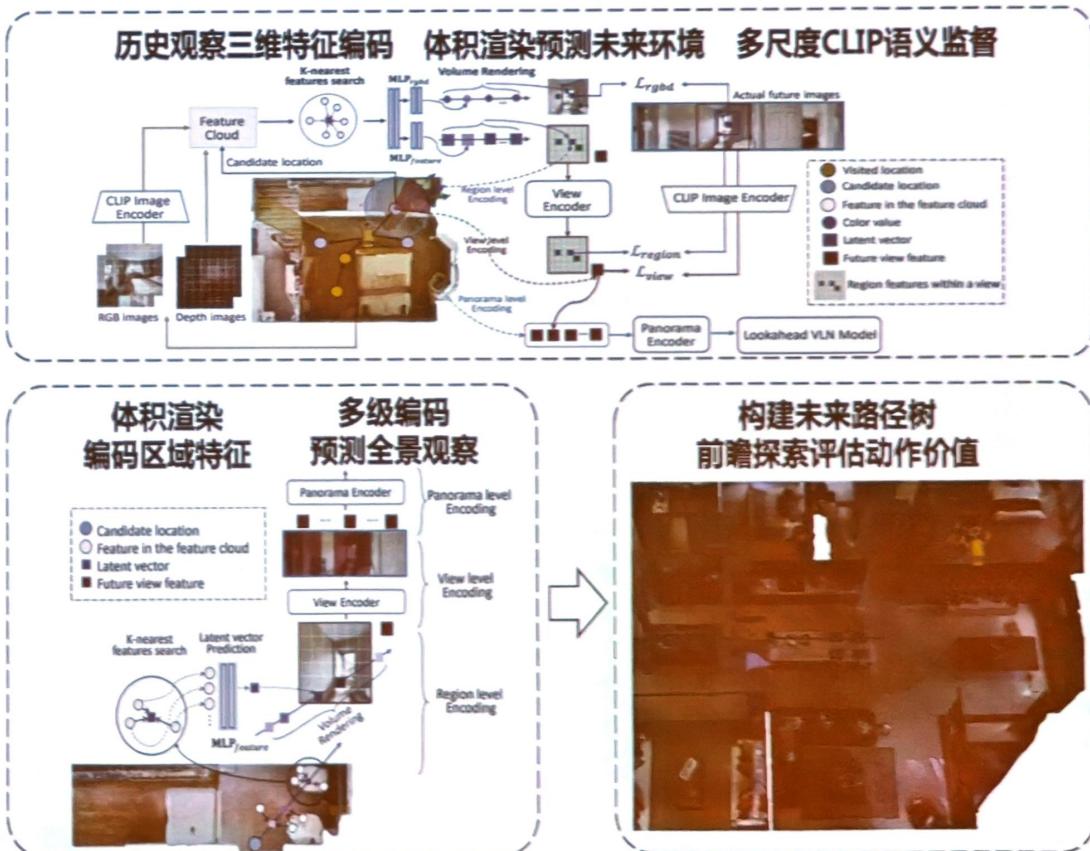
- 现有导航方法采用单步动作预测，  
缺乏长远动作规划能力；传统神  
经辐射场速度慢、泛化能力差，  
难以用于具身智能任务

### ➤ 研究思路

- 在三维特征空间中层次化地进行  
神经辐射表征编码，预测未来环  
境；构建可导航的未来路径树，  
前瞻探索评估各分支的行动价值

### ➤ 效果

- 在R2R-CE, RxR-CE上达到当前  
的最佳性能



# 基于神经辐射表征的前瞻探索策略

## 三维特征云编码与区域表征预测

**二维视觉观察编码为三维特征云**

记录区块特征、三维坐标、观察方向、区块范围



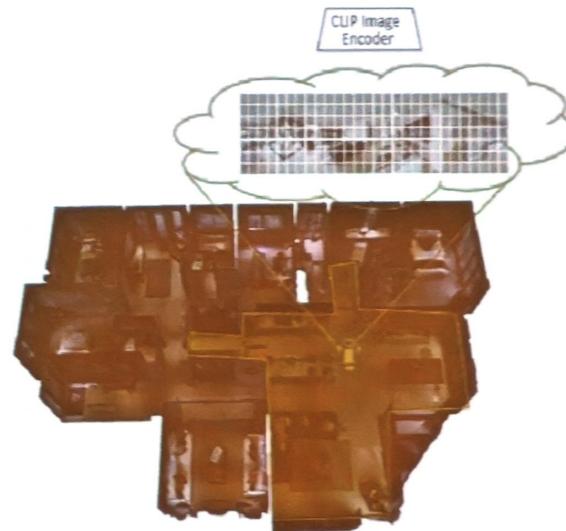
**KD-Tree最邻近特征搜索**

搜索观察射线上采样点的K个最邻近特征



**体积渲染预测区块表征**

聚合K个最邻近特征预测采样点表征和体积密度

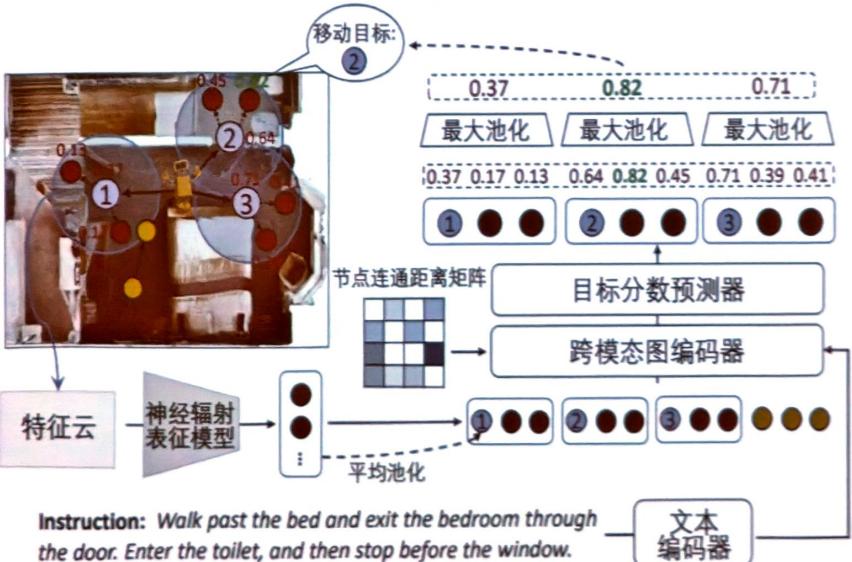


# 基于神经辐射表征的前瞻探索策略

## 构建未来路径树与前瞻探索

### 前瞻导航模型

构建未来路径树，评估长远行动价值



### 导航演示

**Instruction:** Exit the bedroom. Walk the opposite way of the picture hanging on the wall through the kitchen. Turn right at the long white countertop. Stop when you get past the two chairs.

**Observation:**



# 基于神经辐射表征的前瞻探索策略

## 实验评估（与当前方法比较）

表1. R2R-CE数据集上的性能比较

Methods	Val Seen					Val Unseen					Test Unseen				
	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑
CM <sup>2</sup> Georgakis et al. (2022)	12.05	6.10	50.7	42.9	34.8	11.54	7.02	41.5	34.3	27.6	13.9	7.7	39	31	24
WS-MGMap Chen et al. (2022a)	10.12	5.65	51.7	46.9	43.4	10.00	6.28	47.6	38.9	34.3	12.30	7.11	45	35	28
Sim-2-Sim Krantz and Lee (2022)	11.18	4.67	61	52	44	10.69	6.07	52	43	36	11.43	6.17	52	44	37
CWP-RecBERT Hong et al. (2022)	12.50	5.02	59	50	44	12.23	5.74	53	44	39	13.31	5.89	51	42	36
GridMM Wang et al. (2023d)	12.69	4.21	69	59	51	13.36	5.11	61	49	41	13.31	5.64	56	46	39
Reborn An et al. (2022)	10.29	4.34	67	59	56	10.06	5.40	57	50	46	11.47	5.55	57	49	45
Ego <sup>2</sup> -Map Hong et al. (2023)	-	-	-	-	-	-	4.94	-	52	46	13.05	5.54	56	47	41
DREAMWALKER Wang et al. (2023a)	11.6	4.09	66	59	48	11.3	5.53	59	49	44	11.8	5.48	57	49	44
ScaleVLN Wang et al. (2023c)	-	-	-	-	-	-	4.80	-	55	51	-	5.11	-	55	50
BEVBert An et al. (2023a)	-	-	-	-	-	-	4.57	67	59	50	-	4.70	67	59	50
ETPNNav An et al. (2023b)	11.78	3.95	72	66	59	11.99	4.71	65	57	49	12.87	5.12	63	55	48
HNR (Ours)	11.79	3.67	76	69	61	12.64	4.42	67	61	51	13.03	4.81	67	58	50

表2. RxR-CE数据集上的性能比较

Methods	Val Seen					Val Unseen					Test Unseen				
	NE↓	SR↑	SPL↑	NDTW↑	SDTW↑	NE↓	SR↑	SPL↑	NDTW↑	SDTW↑	NE↓	SR↑	SPL↑	NDTW↑	SDTW↑
CWP-CMA Hong et al. (2022)	-	-	-	-	-	8.76	26.59	22.16	47.05	-	10.40	24.08	19.07	37.39	18.65
CWP-RecBERT Hong et al. (2022)	-	-	-	-	-	8.98	27.08	22.65	46.71	-	10.40	24.85	19.61	37.30	19.05
Reborn An et al. (2022)	5.69	52.43	45.46	66.27	44.47	5.98	48.60	42.05	63.35	41.82	7.10	45.82	38.82	55.43	38.42
ETPNav An et al. (2023b)	5.03	61.46	50.83	66.41	51.28	5.64	54.79	44.89	61.90	45.33	6.99	51.21	39.86	54.11	41.30
HNR (Ours)	4.85	63.72	53.17	68.81	52.78	5.51	56.39	46.73	63.56	47.24	6.81	53.22	41.14	55.61	42.89

在连续环境视觉语言导航数据集R2R-CE, RxR-CE 上获得当前的**最佳性能**

# 基于神经辐射表征的前瞻探索策略

## 实验评估（性能与复杂度评估）

图1. 不同未来视角预测方法的预测质量  
(与真实图像特征的余弦相似度)

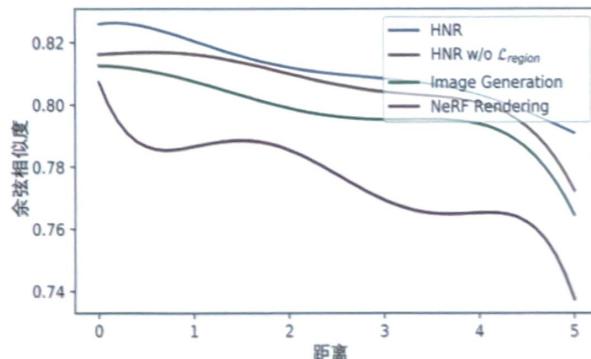


表4. 不同未来视角预测方法的耗时分析

NeRF Rendering	Image Generation	HNR
21.6 Hz (46.3 ms)	12.6 Hz (79.4 ms)	87.3 Hz (11.5 ms)

表3. 不同未来视角预测方法进行前瞻探索的导航性能

#	表征方法	NE↓	OSR↑	SR↑	SPL↑
1	Single View	4.71	64.71	57.21	49.15
2	NeRF Rendering	4.79	65.14	56.55	48.61
3	Image Generation	4.68	66.01	58.35	50.96
4	HNR	4.42	67.48	60.74	51.27
5	Ground truth	4.13	71.29	63.13	54.59

- Single View: 不使用前瞻探索的Baseline方法
- NeRF Rendering: 使用NeRF方法预测未来视角RGB图像
- Image Generation: 使用生成模型预测未来视角RGB图像
- HNR: 我们的层次化神经辐射表征模型
- Ground truth: 真实的未来视角图像，反映了性能上界

在导航场景下新视角预测的**质量和速度**均具备显著优势

# 基于三维特征场和语义通行地图的Sim-to-Real导航

## 研究背景与动机

### ➤ 挑战与动机

- 口 绝大多数机器人仅配备窄视角的单目相机 ( $FOV < 90^\circ$ )，受限的视野极大限制了视觉语言导航性能

### ➤ 研究思路

- 口 预测可通行地图，获得全景环境的可通行性感知能力；
- 口 构建3D特征场，预测相机视野外的环境表征，获得全景语义理解能力

### ➤ 效果

- 口 达到单目相机设置下R2R-CE, RxR-CE数据集的最佳性能，导航成功率领先大于7%
- 口 兼容主流的视觉语言导航模型，提供了高性能的通用Sim-to-real导航方案

*Instruction: Go past the table with a laptop, find the basketball.*



使用单目相机的视觉语言导航模型

- 受限的视场角 $< 90^\circ$
- 贫弱的导航成功率 $< 39\%$
- 兼容绝大多数机器人

使用全景相机的视觉语言导航模型

- 辽阔的视野 $\approx 360^\circ$
- 优异的导航成功率 $> 57\%$
- 极少量机器人配备全景RGB-D相机

兼备？

- 兼容绝大多数机器人
- 优异的导航成功率

# 基于三维特征场和语义通行地图的Sim-to-Real导航

## 视觉语言导航的Sim-to-Real挑战与思路

### ➤ 挑战

- 模拟器环境与现实环境的视觉图像差异巨大
- 模拟器环境的物理反馈(碰撞、摩擦)不真实

- 实体机器人受限的感知能力(障碍物感知、场景语义感知)

- 实体机器人的传感器误差(定位不准、相机失真)

- 现实环境导航需要低延时、低计算开销

- 现实环境布局更复杂，导航规划难度高

- 训练数据欠缺，模型泛化能力差

- 训练数据的语言指令呆板，难以实际应用

### ➤ 解决策略

- 构建语义通行地图实现避障与导航点预测
- 构建三维特征场实现全景环境语义理解

- 摒弃原子动作预测(左/右转15°,前进0.25米)
- 直接进行高层规划(推理下一步导航点)

- 构建语义通行地图与拓扑路径地图，实现长期路径规划和纠错

### ➤ 效果

兼容性

无需全景相机或激光雷达，  
兼容大多数VLN模型实机部署

实时性

GTX 1060单卡实时导航

导航性能

模拟器与现实环境导航  
成功率提升均超7%

# 基于三维特征场和语义通行地图的Sim-to-Real导航

## 基于三维特征场的环境感知

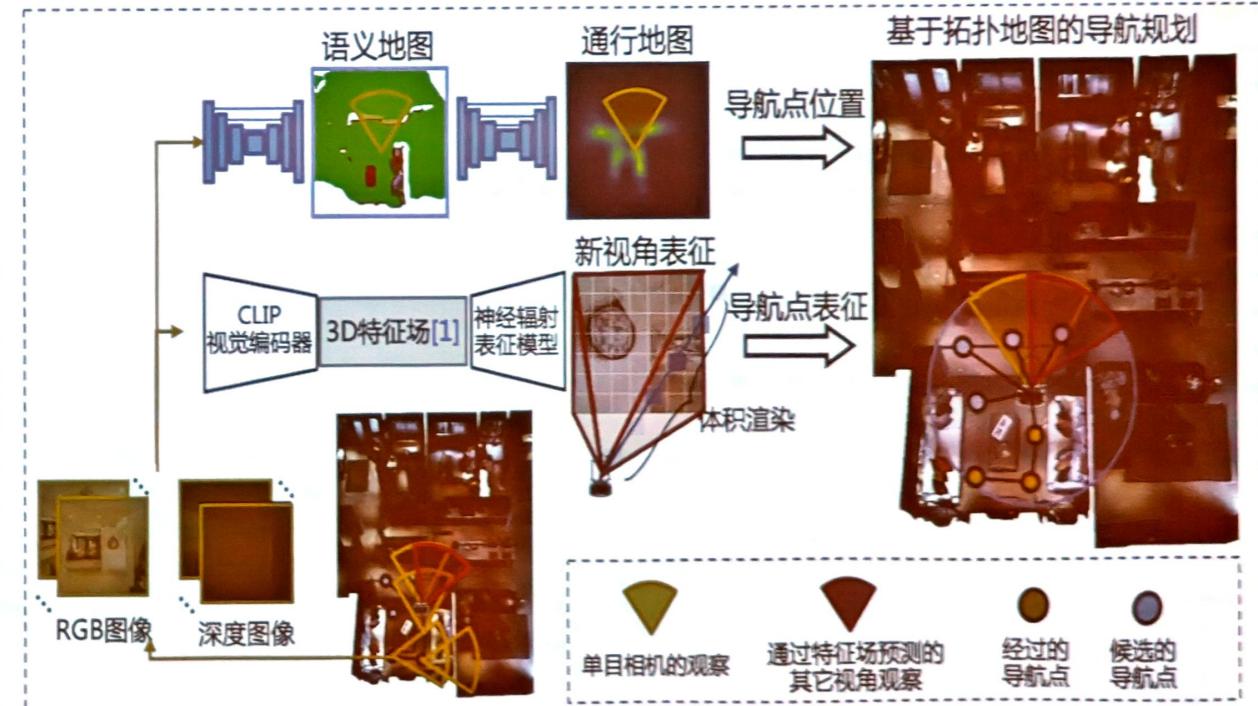
构建语义地图，预测可通行区域和可导航点位置



编码视觉特征到**三维的特征场**空间，解码预测可导航点表征



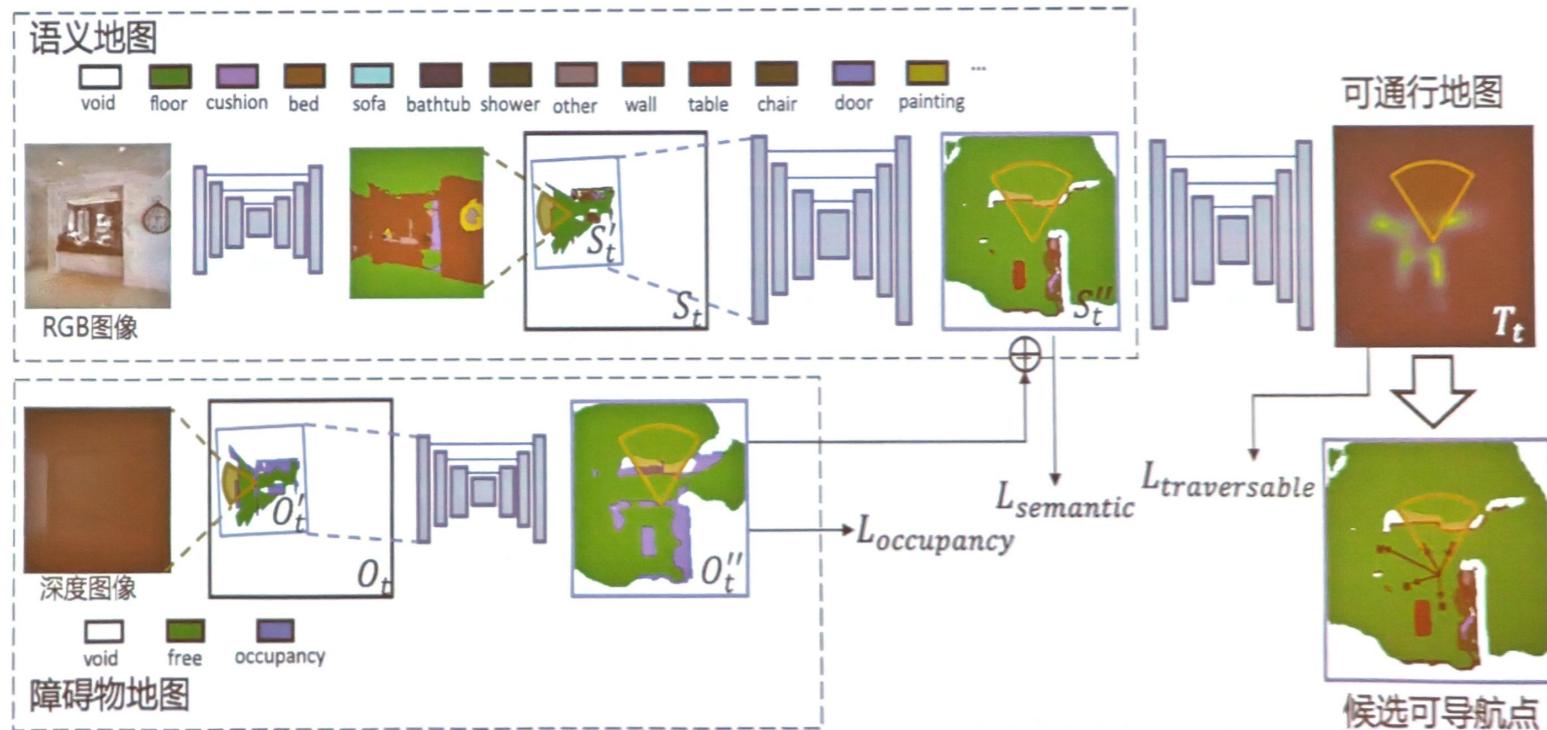
基于可导航点位置与表征构建全局拓扑地图，进行导航规划



[1] Lookahead Exploration with Neural Radiance Representation for Continuous Vision-Language Navigation. CVPR 24 Highlight. Our paper.

# 基于三维特征场和语义通行地图的Sim-to-Real导航

## 基于语义通行地图的导航点预测



# 基于三维特征场和语义通行地图的Sim-to-Real导航

## 实验评估

表1. 模拟器环境R2R-CE数据集上的性能比较 ( SR为导航成功率 )

Camera	Methods	Val Unseen			Test Unseen				
		NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑
Panoramic	Sim-2-Sim (Krantz and Lee, 2022)	6.07	52	43	36	6.17	52	44	37
	CWP-CMA (Hong et al., 2022)	6.20	52	41	36	6.30	49	38	33
	CWP-RecBERT (Hong et al., 2022)	5.74	53	44	39	5.89	51	42	36
	GridMM (Wang et al., 2023)	5.11	61	49	41	5.64	56	46	39
	BEVBert (An et al., 2023)	4.57	67	59	50	4.70	67	59	50
	ETPNav (An et al., 2024)	4.71	65	57	49	5.12	63	55	48
Monocular	CM <sup>2</sup> (Georgakis et al., 2022)	7.02	41.5	34.3	27.6	7.7	39	31	24
	WS-MGMap (Chen et al., 2022)	6.28	47.6	38.9	34.3	7.11	45	35	28
	NaVid (Zhang et al., 2024)	5.47	49.1	37.4	35.9	-	-	-	-
Monocular	GridMM w/ Feature Fields	6.36	52.7	40.3	28.7	6.86	49.4	37.5	25.5
	ETPNav w/ Feature Fields	5.95	55.8	44.9	30.4	6.24	54.4	43.7	28.9

表2. 模拟器环境RxR-CE数据集上的性能比较 ( SR为导航成功率 )

Camera	Methods	Val Unseen			
		NE↓	OSR↑	SR↑	SPL↑
Panoramic	CWP-CMA (Hong et al., 2022)	8.76	-	26.6	22.2
	CWP-RecBERT (Hong et al., 2022)	8.98	-	27.1	22.7
	ETPNav (An et al., 2024)	5.6	-	54.8	44.9
Monocular	CM <sup>2</sup> (Georgakis et al., 2022)	8.98	25.3	14.4	9.2
	WS-MGMap (Chen et al., 2022)	9.83	29.8	15.0	12.1
	A <sup>2</sup> Nav (Chen et al., 2023)	-	-	16.8	6.3
Monocular	NaVid (Zhang et al., 2024)	8.41	34.5	23.8	21.2
	ETPNav w/ Feature Fields	8.79	36.7	25.5	18.1

表3. LocoBot机器人真实环境的性能比较

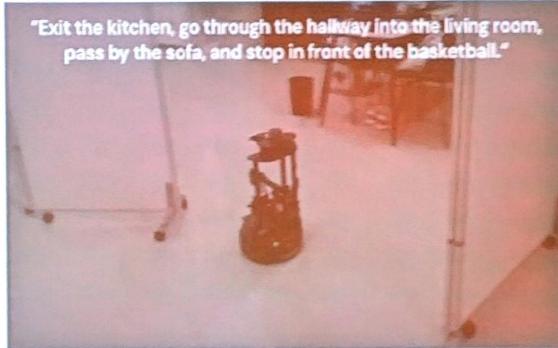
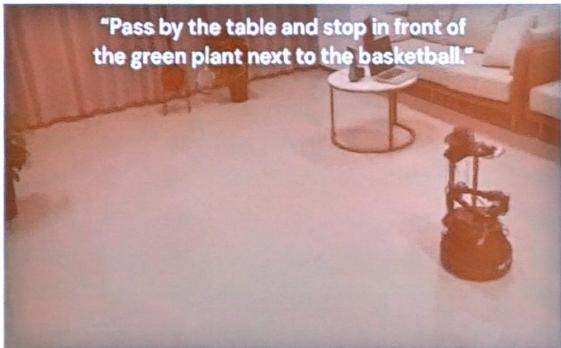
Methods	Living Room		Bedroom		Kitchen		Bathroom		All	
	OSR↑	SR↑	OSR↑	SR↑	OSR↑	SR↑	OSR↑	SR↑	OSR↑	SR↑
CM <sup>2</sup> (Georgakis et al., 2022)	17.1	11.4	23.5	11.8	12.9	6.5	17.6	11.8	17.0	10.0
WS-MGMap (Chen et al., 2022)	22.9	14.3	29.4	23.5	22.6	12.9	17.6	11.8	23.0	15.0
GridMM w/ Feature Fields	51.4	40.0	41.2	29.4	48.4	41.9	41.2	23.5	47.0	36.0
ETPNav w/ Feature Fields	54.3	45.7	52.9	47.1	58.1	38.7	47.1	35.3	54.0	42.0

在模拟器环境R2R-CE, RxR-CE 数据集单目相机下获得最佳导航成功率，现实环境导航性能大幅提升

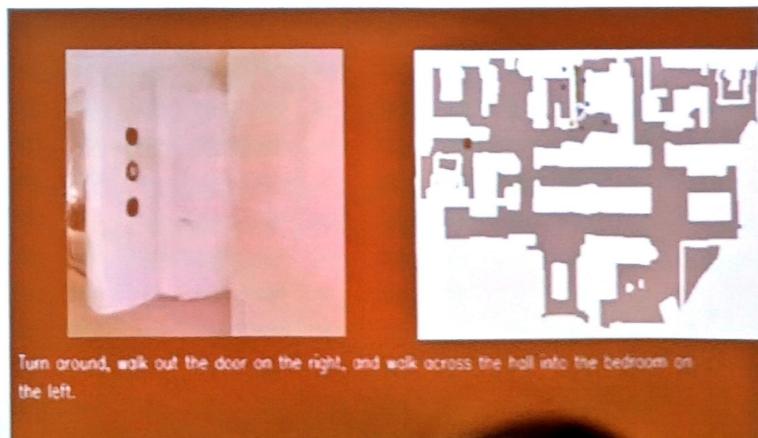
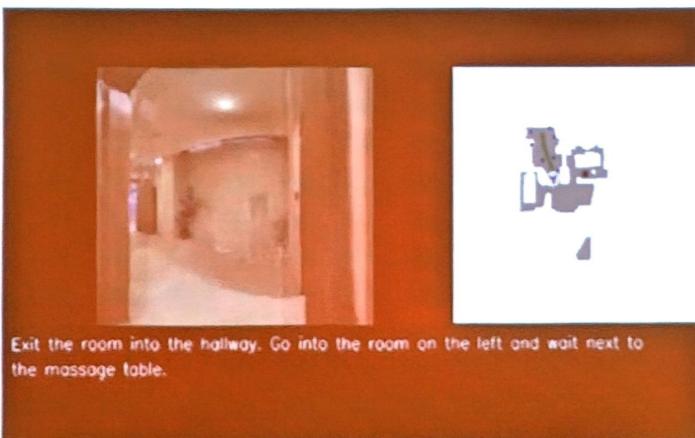
# 基于三维特征场和语义通行地图的Sim-to-Real导航

真实环境

## 视觉语言导航演示



模拟器环境



# 边缘计算需求：Locobot机器人视觉导航

实验平台：Locobot  
开源机器人  
传感器：

- RGB-D
- 激光雷达
- 里程计

I3CPU



	SLAM建图	目标检测	动作规划
边缘处理器算力 I3-CPU	2x3.0G Hz	2x3.0G Hz	2x3.0G Hz
效率	8k 点云/s 1m/s	2FPS	2次/s
上位机算力- RTX3060-GPU	-	3584x1.32G Hz	3584x1.32G Hz
效率	-	15FPS	15次/s

算力瓶颈：深度学习模型vs低功率CPU



提升途径：深度学习芯片/加速棒

# 汇报大纲

具身智能介绍

具身导航研究进展

虚拟到现实环境的演示

总结与展望

# 系统演示：Locobot机器人视觉导航



视觉导航：动态环境、实例化、多目标、可交互

实验平台：Locobot  
开源机器人  
传感器：  
➤ RGB-D  
➤ 激光雷达  
I3CPU



真实环境

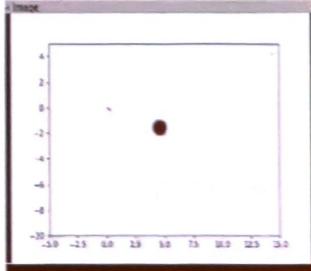


多样化环境布局



边导航边建图

- 支持动态环境、动态目标导航
- 支持在新环境布局无地图冷启动视觉导航
- 支持边导航边建图，导航能力持续增长



# 系统演示：Locobot机器人视觉导航

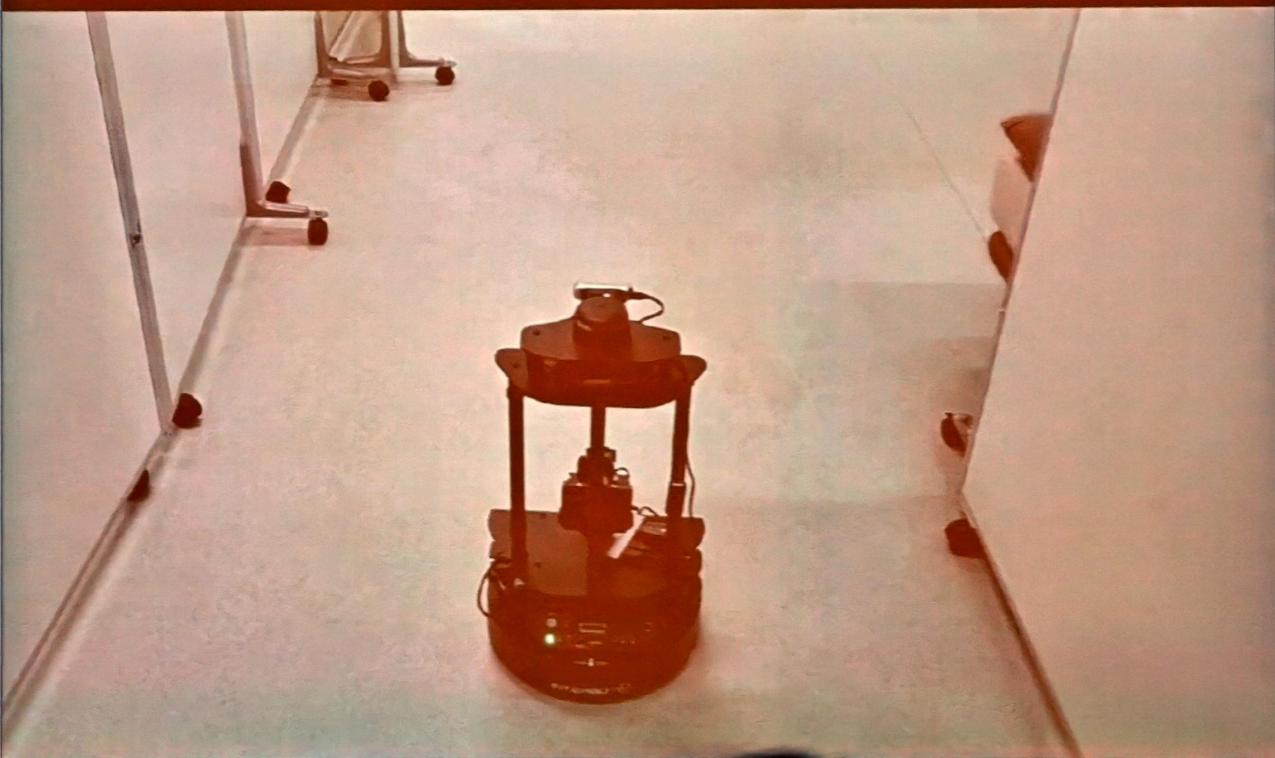
目标导航-场景图规划

任务：  
去卧室找一下闹钟

拓扑图规划：  
去卧室

执行：  
正在前往卧室

## 去卧室找一下闹钟



# 系统演示：Locobot机器人视觉导航

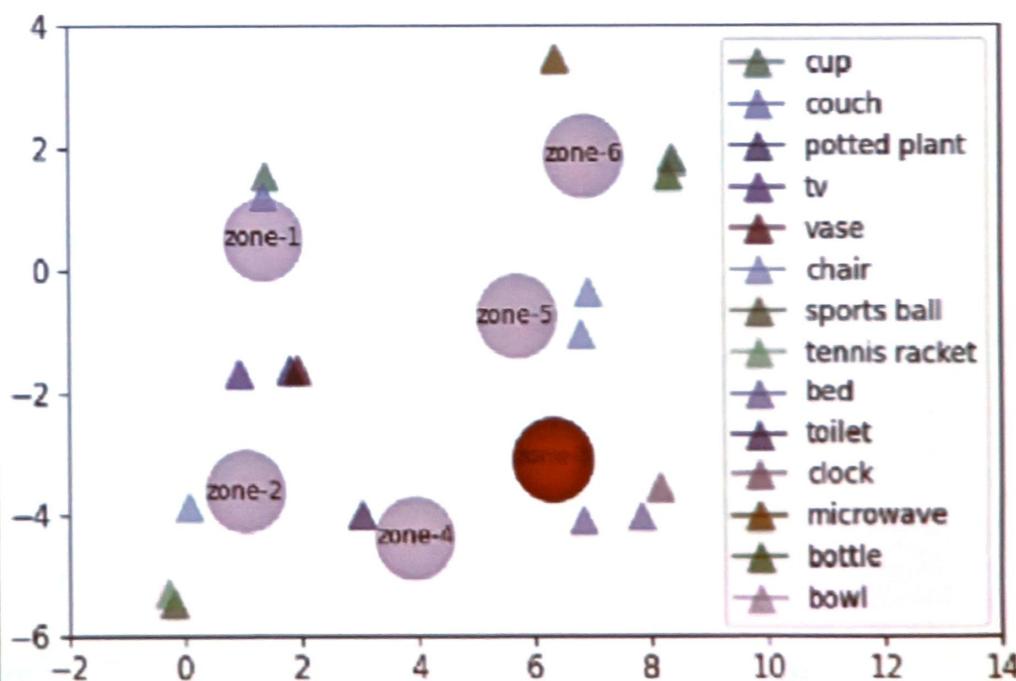
目标导航-场景图规划

任务：  
去卧室找一下闹钟

拓扑图规划：  
去卧室

执行：  
正在前往卧室

## 语义拓扑图规划



# 系统演示：Locobot机器人视觉导航

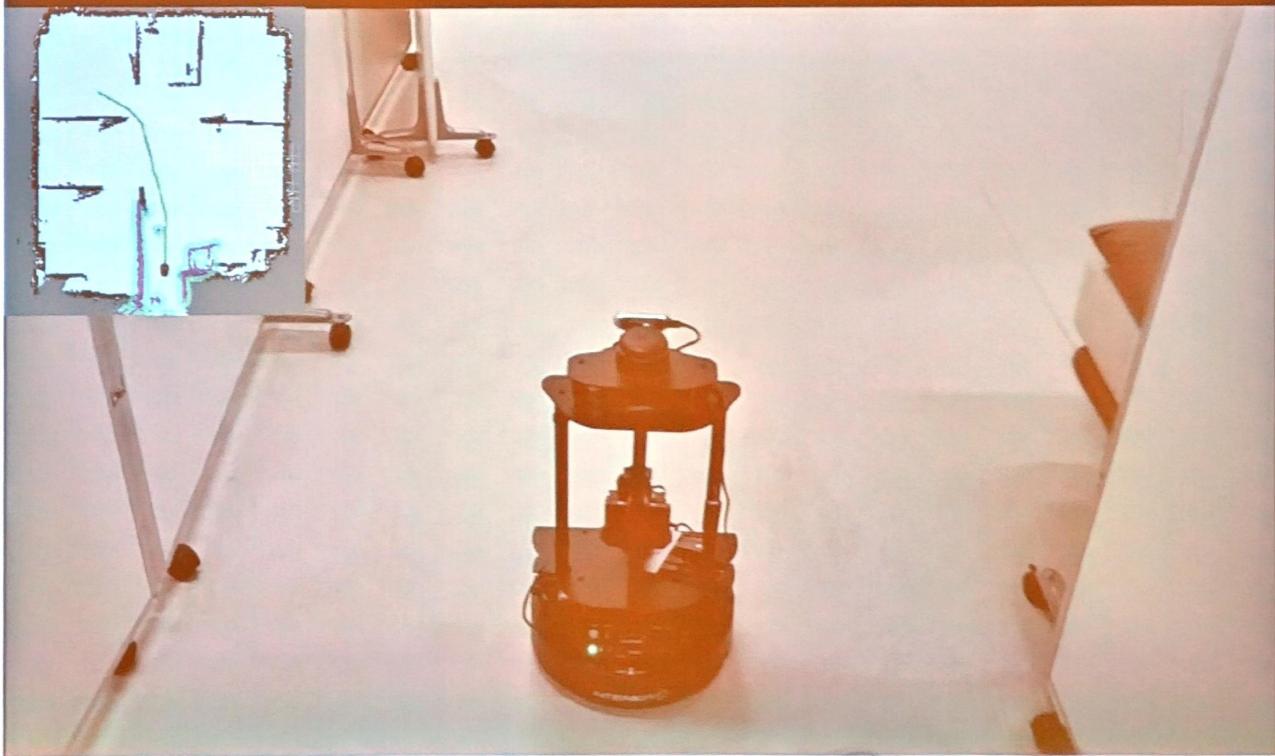
目标导航-场景图规划

任务：  
去卧室找一下闹钟

拓扑图规划：  
去卧室

执行：  
正在前往卧室

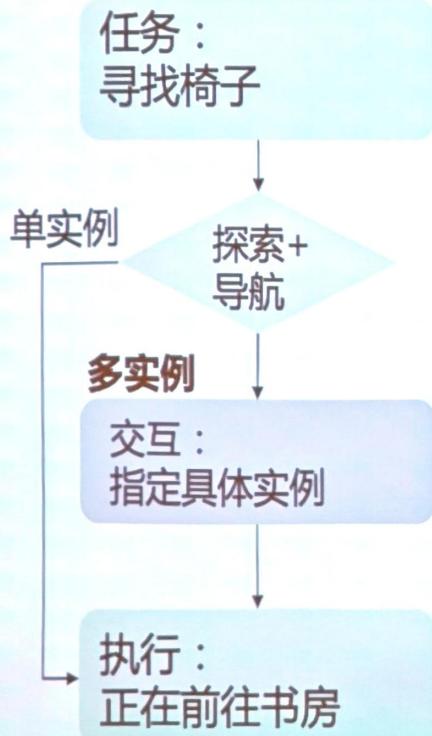
## 正在前往卧室



# 系统演示：Locobot机器人视觉导航

实例级目标导航-交互

## 在哪里可以找到椅子



# 系统演示：Locobot机器人视觉导航

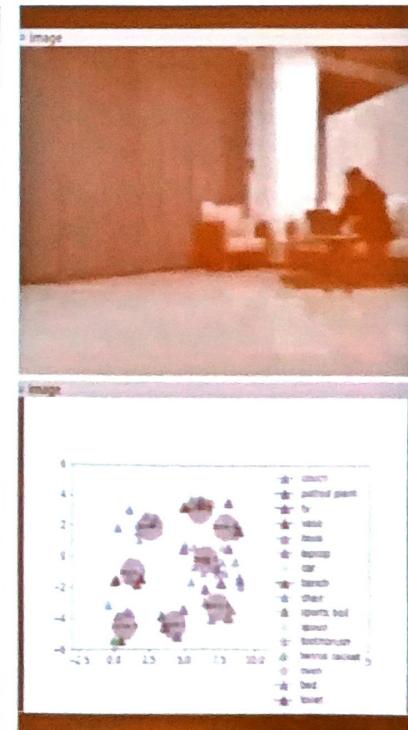


任务：把郭鑫龙的闹钟给周龙

- 1 ) 导航到郭鑫龙；2 ) 放置闹钟；3 ) 导航到周龙；4 ) 拿取闹钟



第三人称视角



第一人称视角

# 项目基础



2011年

机器博士：  
机器专家  
关键技术



2015年

北京市科技  
计划：家庭  
陪伴智能机  
器人（联想）



2020年

AI2030：人机协  
同智能软硬件技  
术研究（外骨骼）



2021年

华为合作：  
家庭环境场  
景理解与拓  
扑导航



2023年

基金委联合  
重点项目：  
野外环境场  
景感知

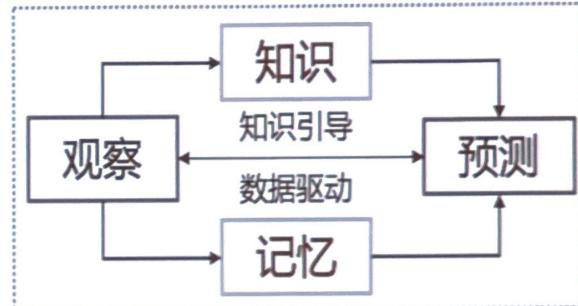


# 总结

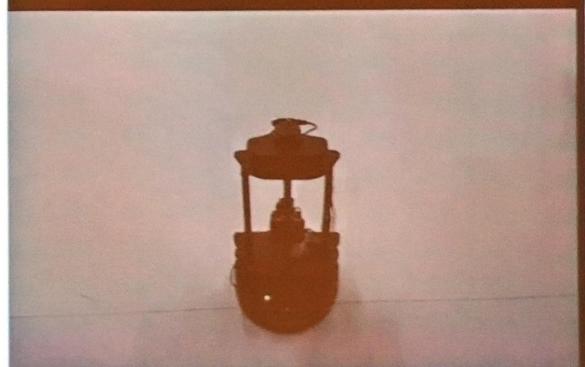
- 具身导航是具身智能的重要任务之一，在模拟器上的模型已有较高的精度
- 面向实体机器人和真实环境的具身智能研究有广阔应用价值

- 提出了知识增强具身导航框架
- 创新点包括：多级场景图构建与在线更新、场景知识因果分析、知识推理与增强、场景知识网格记忆

- 虚拟到现实：在真实环境的导航准确率逐步接近模拟器



在哪里可以找到椅子



# 成果总结

## 相关论文列表-1

- [1-IJCAI2017] Shuqiang Jiang, Weiqing Min, Xue Li, Huayang Wang, Jian Sun, Jiaqi Zhou. Dual Track Multimodal Automatic Learning through Human-Robot Interaction. IJCAI 2017, pp. 4485-4491
- [2-ICCV21] Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, Shuqiang Jiang. Hierarchical Object-to-Zone Graph for Object Navigation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15130-15140
- [3-CVPR23] Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, Shuqiang Jiang: Layout-based Causal Inference for Object Navigation, IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2023
- [4-CVPR23] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, Shuqiang Jiang; KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2583-2592
- [5-ICCV23] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Shuqiang Jiang. GridMM: Grid Memory Map for Vision-and-Language Navigation, ICCV 23
- [6-NeurIPS23] Xiaohan Wang, Yuehu Liu, Xinhang Song, Beibei Wang, Shuqiang Jiang. CaMP: Causal Multi-policy Planning for Interactive Navigation in Multi-room Scenes. NeurIPS 2023.
- [7-TIP23] Haitao Zeng, Xinhang Song, Shuqiang Jiang, Multi-Object Navigation using Potential Target Position Policy Function, IEEE TIP 2023
- [8-ECCV22] Sixian Zhang, Weijie Li, Xinhang Song, Yubing Bai, Shuqiang Jiang: Generative Meta-Adversarial Network for Unseen Object Navigation. ECCV (39) 2022: 301-320
- [9-ACM MM 21] Weijie Li, Xinhang Song, Yubing Bai, Sixian Zhang, Shuqiang Jiang: ION: Instance-level Object Navigation. ACM Multimedia 2021: 4343-4352

# 合作者



宋新航 副研究员  
多模态感知与视觉导航



张思贤 博士生  
视觉目标导航



曾海涛 博士生  
多目标视觉导航



黎维婕 硕士生  
实例级目标导航



白宇兵 硕士生  
视觉导航-虚拟到现实



黎向阳 助理研究员  
视觉语言导航 ( VLN )



杨嘉豪 博士生  
记忆增强VLN



刘烨琦 博士生  
连续环境下VLN



王子涵 硕士生  
离散环境下VLN



郭皓华 硕士生  
视觉语言定位

联系方式 : [sqjiang@ict.ac.cn](mailto:sqjiang@ict.ac.cn)