具身智能-09

刘华平

2025年4月16日

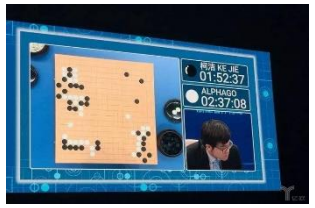
课程内容安排

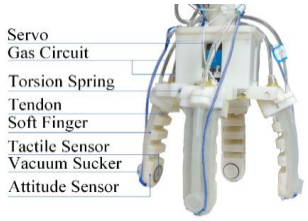
课次	周次	上课内容	软件
1	1	绪论	
2	2	深度学习	
3	3	强化学习1	Gym, Mujuco
4	4	强化学习2	Gym, Mujuco
	5	作业准备	
5	6	自监督与持续学习	
	7	开题	Powerpoint
6	8	形态智能	Gym, Mujuco
7	9	视觉导航：VLN	AI2THOR
8	10	主动感知：VSN，EQA	AI2THOR
	11	五一放假	
9	12	具身学习	AI2THOR
10	13	多体智能	AI2THOR
11	14	面向具身智能的AIGC	AI2THOR
	15-16	成果准备与展示	Powerpoint

具身智能的体系

- 狭义与广义的具身智能

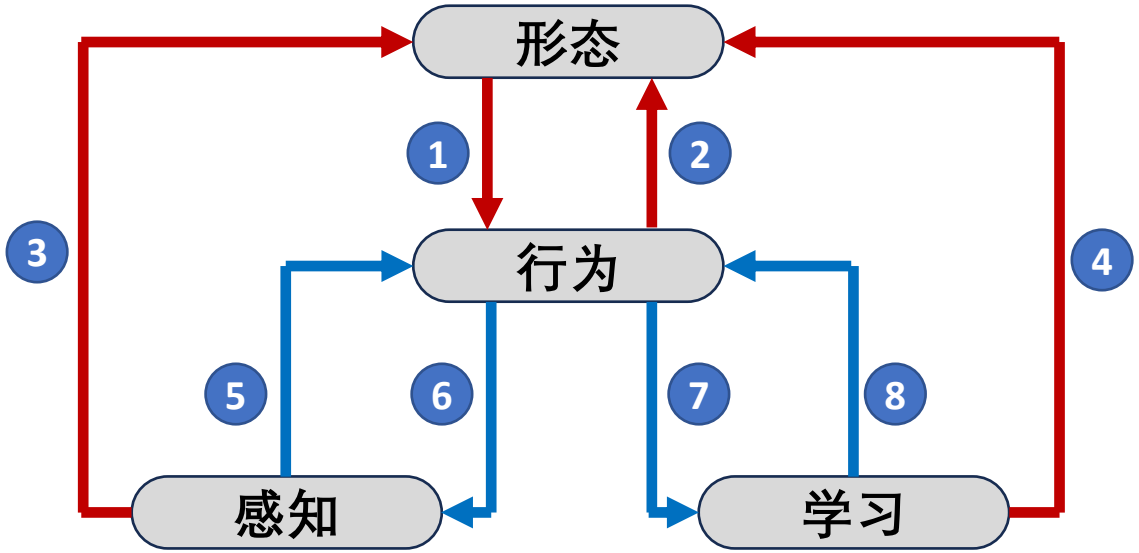
离 身 智 能 具 身 智 能

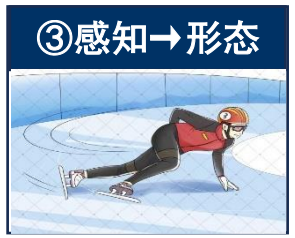




具身智能的体系

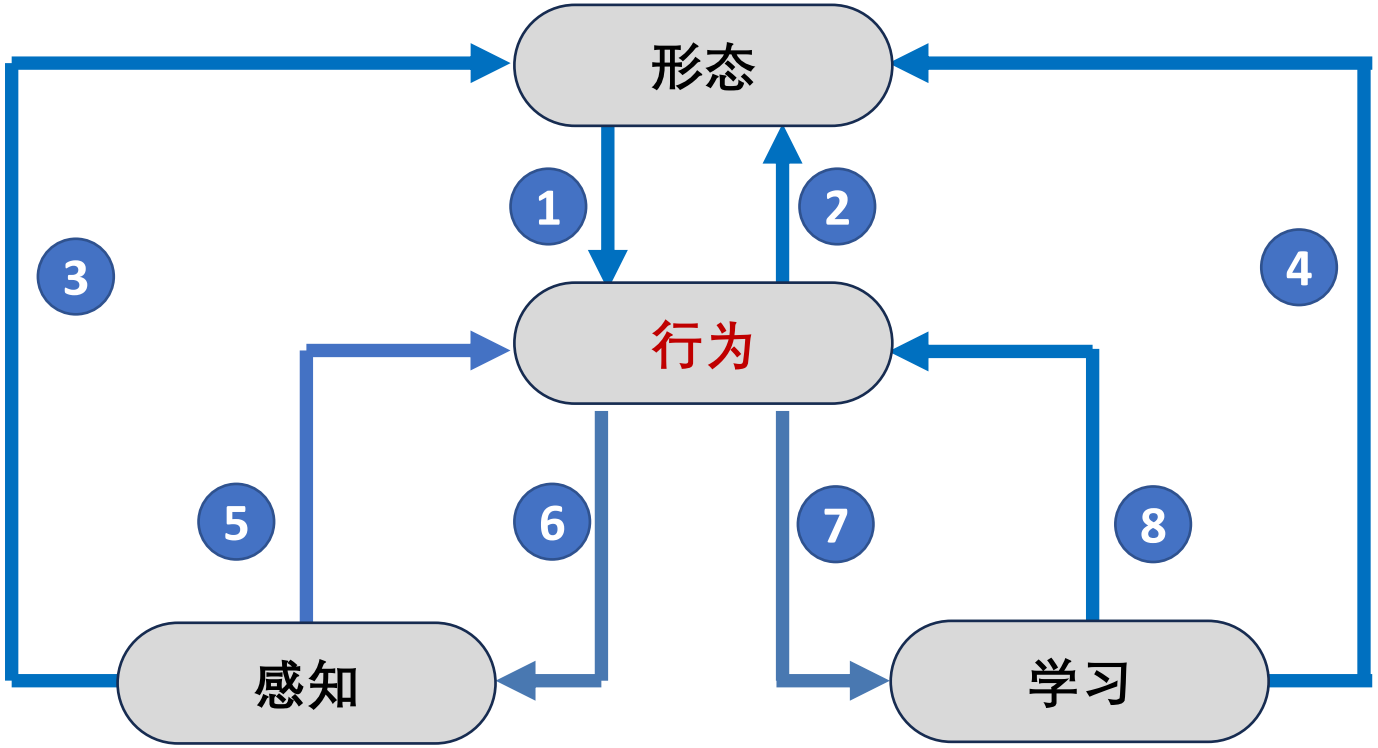
➢ 具身智能的体系结构





① 基于形态的行为生成　③ 基于感知的形态变换　⑤ 基于感知的行为生成　⑦ 基于行为的自主学习

② 基于行为的形态控制　④ 基于学习的形态优化　⑥ 基于行为的主动感知　⑧ 基于学习的行为优化


形态
1
2
3
4
行为
5
6
7
8
感知
学习

具身智能的体系

- 行为

导航

movement

locomotion

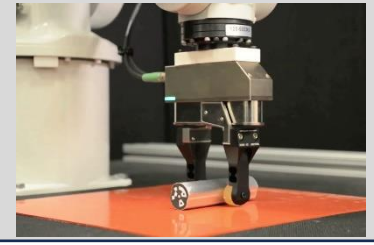
flying

locomotion

操作

manipulation

grasp



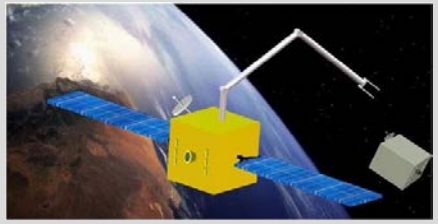
pick&place

assembly

mobile manipulation

flying manipulation

space manipulation



marine manipulation

- 导航
- 视觉导航
- 视觉语言导航

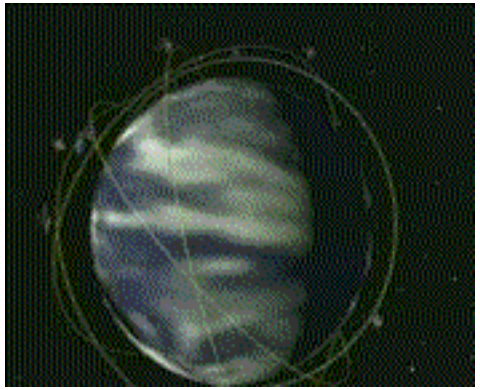
1 导航

Navigation[1] is a field of study that focuses on the process of monitoring and controlling the movement of a craft or vehicle from one place to another.[2] The field of navigation includes four general categories: land navigation,[3] marine navigation, aeronautic navigation, and space navigation.[1]

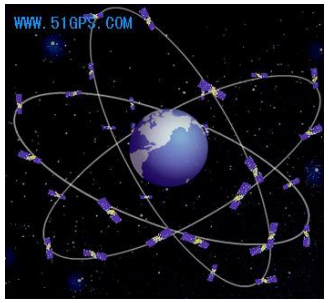
- 我在哪
- 我要去哪
- 我怎么去

1 导航

➢ 定位导航

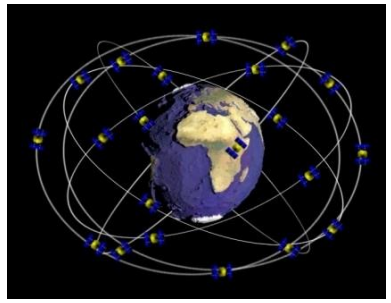


GPS



欧洲 伽利略
（30星）

俄罗斯GLONASS
（24星）

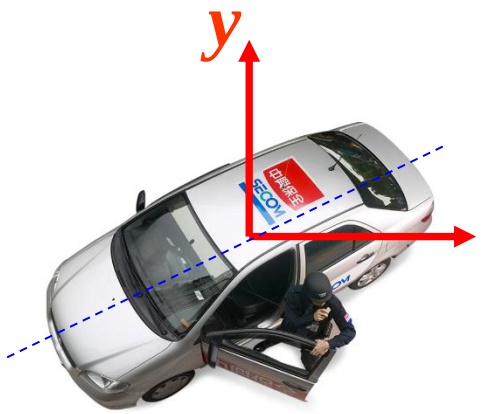


北斗
5颗静止轨道卫星
30颗非静止轨道卫星

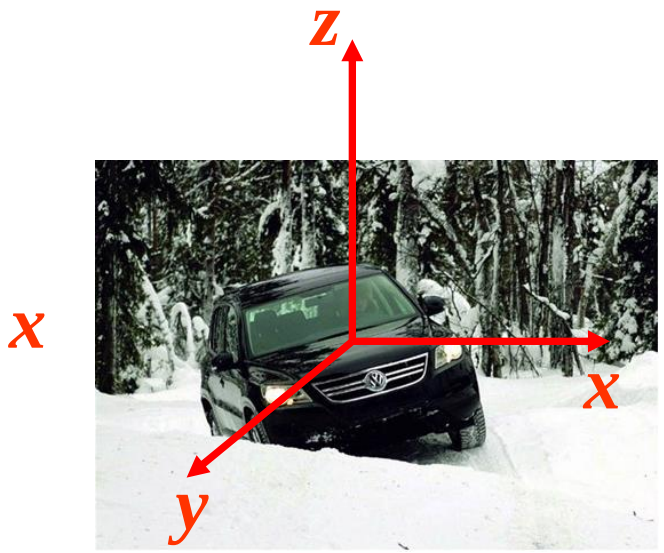
➢ 定位导航

1 导航

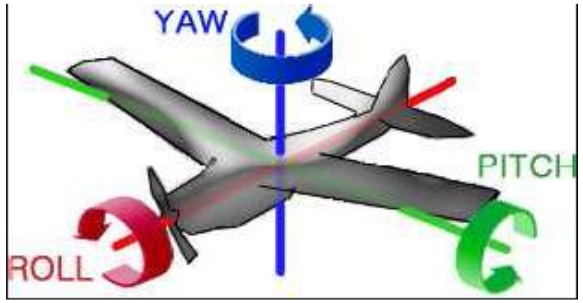
➢ 定位导航



θ



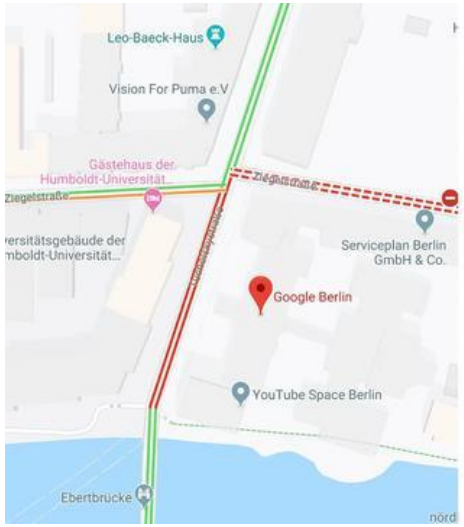
$\theta \quad \varphi \quad \gamma$



横滚，俯仰，偏航

1 导航

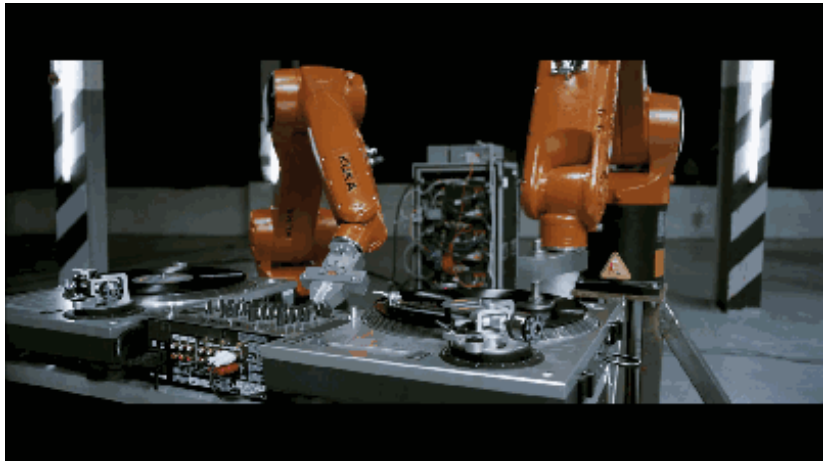
➢ 定位导航


Leo-Baeck-Haus
Vision For Puma e.V
Gästehaus der Humboldt-Universität
Ziegelstraße
Serviceplan Berlin GmbH & Co.
Google Berlin
YouTube Space Berlin
Ebertbrücke


北三环
海淀区
石景山区
丰台区
北京西站
北京南站

1 导航

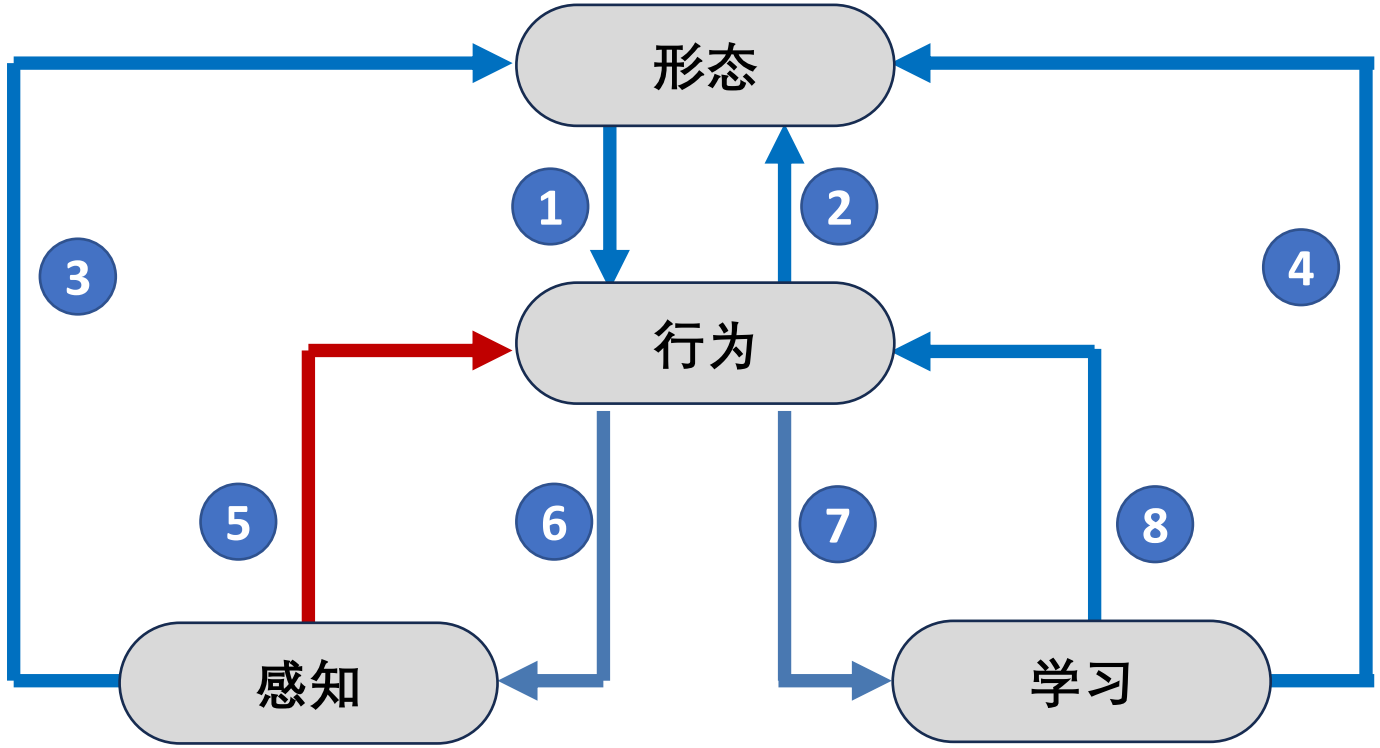
- 孤立的行为



1 导航

- 孤立的行为

2 感知→行为：视觉导航



2 感知→行为：视觉导航

Computer Vision was originally a summer project given to an undergraduate student

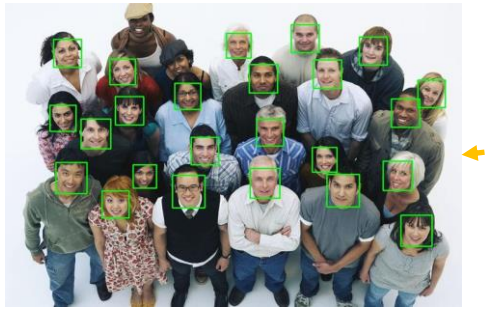
In 1966 Marvin Minsky asked Sussman to “spend the summer linking a camera to a computer and getting the computer to describe what is saw".

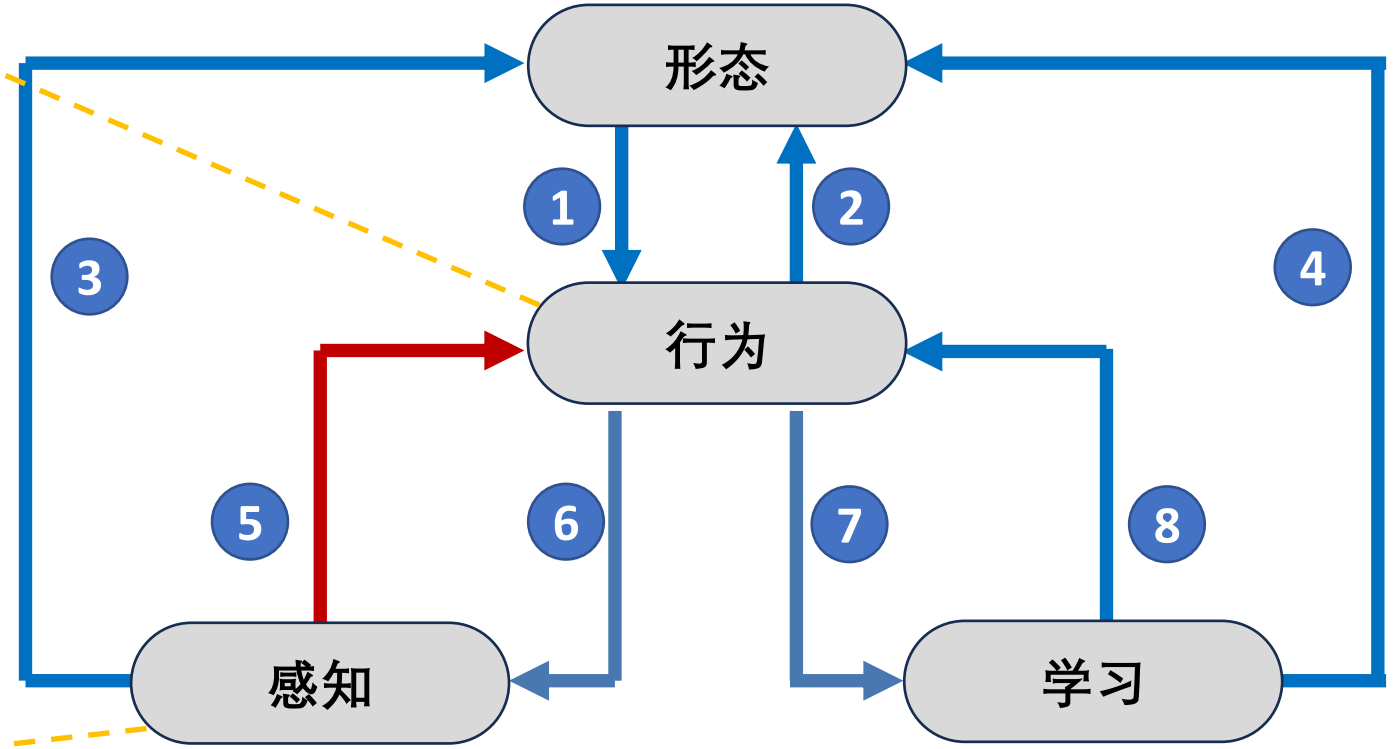
THE NEW YORKER

Fighting fake stories with real ones.

Computer Vision was never a summer project for a single student, nor did it aim to make a complete working vision system. Maybe it was too ambitious for its time, but it's unlikely that the researchers involved thought that it would be completely solved at the end. Finally, Computer Vision as we know it today is vastly different to what it was thought to be in 1966. Today we have many topics derived from CV such as inpainting, novel view generation, gesture recognition, deep learning, etc.

2 感知→行为：视觉导航



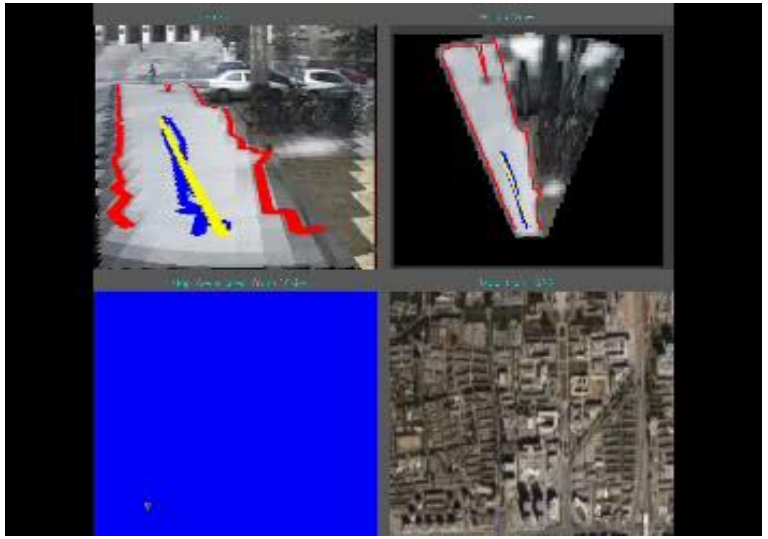


2 感知→行为：视觉导航


Camera
Turn L
Stop
Turn R
FW
BW
Stop Watch
Angle Correcting
Turning Left
Idle
Turning Right
Go Pass
Slow Down
Angle
Distance
QR Angle
QR Code Data


HOT
但研发人员可不打算放过它：走，继续搬砖去！
Boston Dynamics

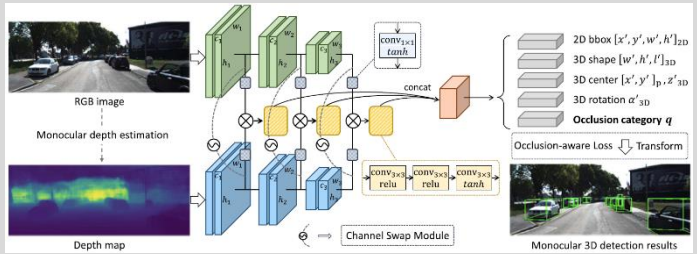
2 感知→行为：视觉导航



2 感知→行为：视觉导航

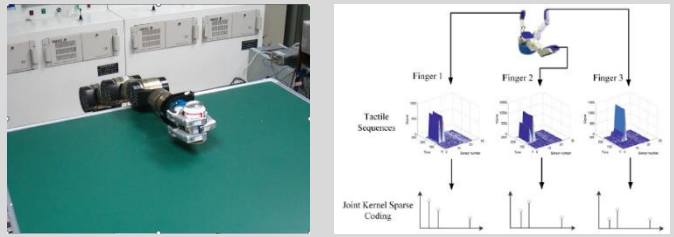
➢ 多模态融合

视觉目标检测



- Fine-grained multi-level fusion for anti-occlusion monocular 3D object detection, IEEE Transactions on Image Processing, 2022.

触觉目标识别

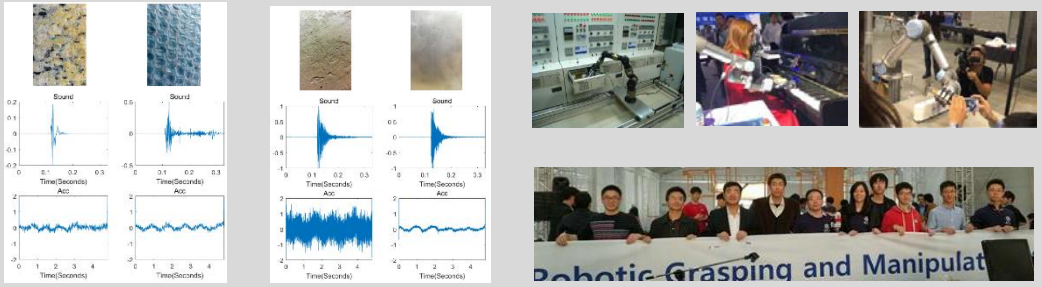


- Robotic Tactile Perception and Understanding: A Sparse Coding Method, Springer, 2018

听觉目标定位

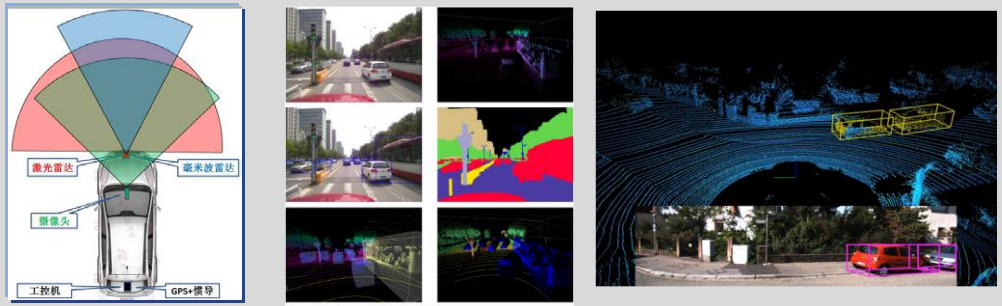
- Robotic room-level localization using multiple sets of sonar measurements, IEEE Transactions on Instrumentation and Measurement, 2017

视-听-触觉融合



- Multi-modal measurements fusion for surface material categorization, IEEE Transactions on Instrumentation and Measurement, 2018

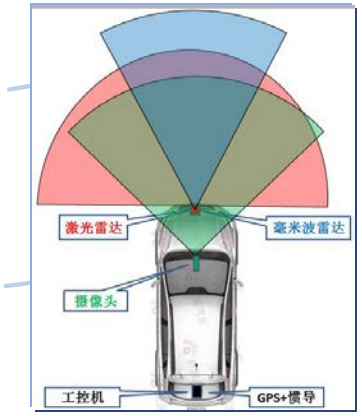
光学-雷达融合



- OpenMPD: An open multimodal perception dataset for autonomous driving, IEEE Transactions on Vehicular Technology, 2021

2 感知→行为：视觉导航

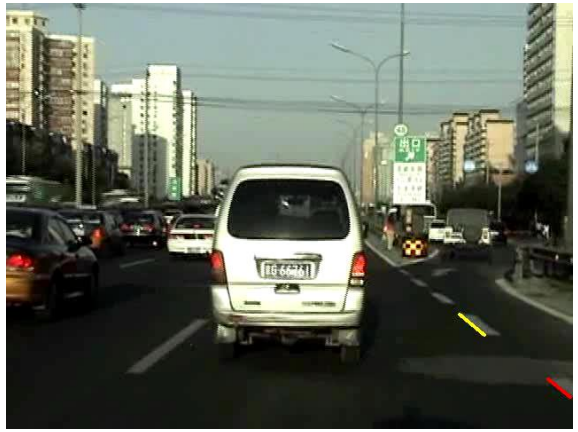
➢ 多模态融合



- Informative data selection with uncertainty for multimodal object detection, IEEE Transactions on Neural Networks and Learning Systems, 2023
- OpenMPD: An open multimodal perception dataset for autonomous driving, IEEE Transactions on Vehicular Technology, 2022

2 感知→行为：视觉导航

- 风险






2 感知→行为：视觉导航

➢ 风险

2016年1月20日，京港澳高速河北邯郸段发生全球首例自动驾驶事故。

2016年5月7日，美国佛罗里达州高速公路发生车祸，驾驶员当场死亡。

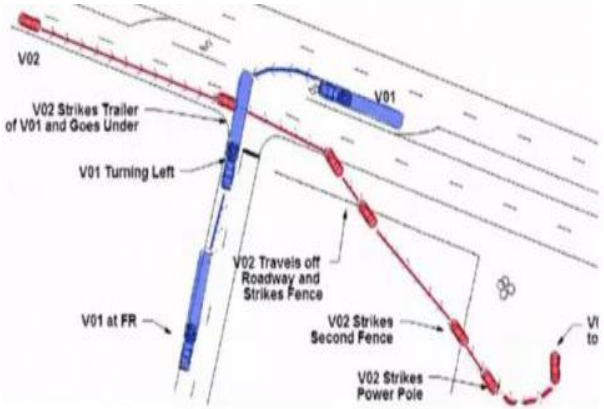
2018年3月19日，美国亚利桑那州Uber路测无人车，撞死了一位横穿马路的妇女。

2018年5月11日，特斯拉在美国犹他州南乔丹以60英里的时速撞上一辆消防车。

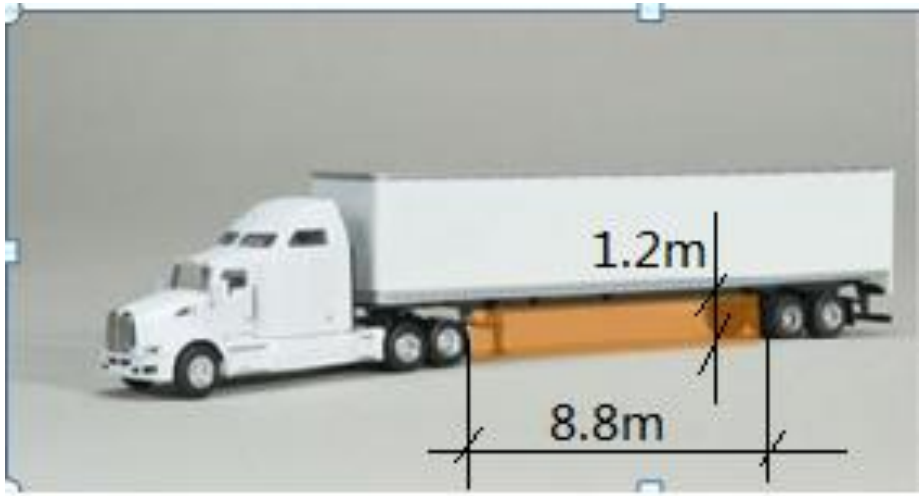
2 感知→行为：视觉导航

➢ 风险

特斯拉Model S



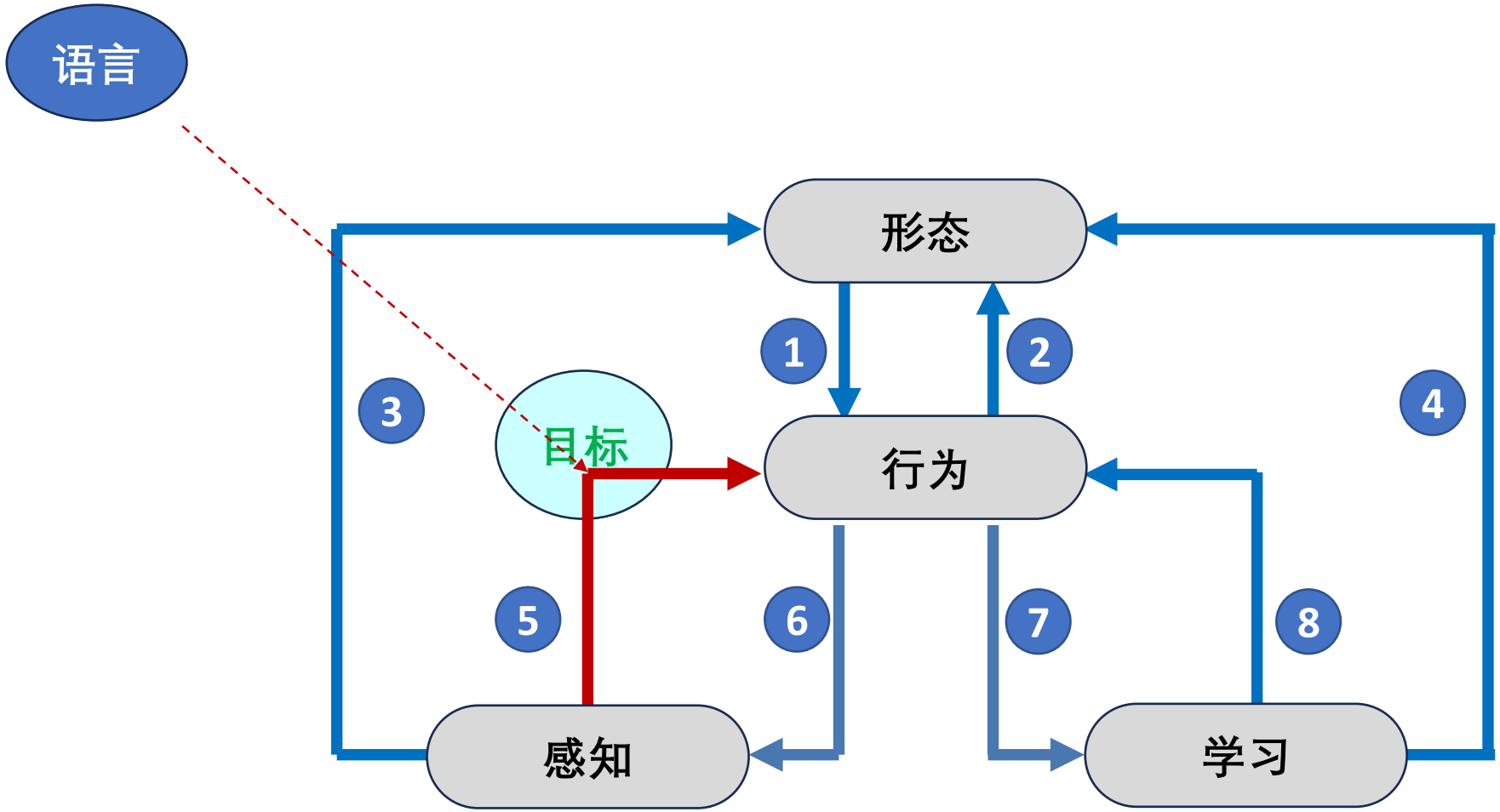
事故发生前后双方驾驶路线图



- 原因一：Tesla 使用的前向视觉感知系统eyeq3只能识别前车尾部，不能识别侧面
- 原因二：雷达可能测算了出了前方有一个巨大的障碍物，误判为悬挂在道路上方的交通指示牌
- 原因三：后融合方案鲁棒性不高。视觉系统认为是一朵白云，雷达认为是一个悬空的交通指示牌

2 感知→行为：视觉导航

3 感知→行为：视觉语言导航



3 感知→行为：视觉语言导航

➢ 语言的重要性

The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity.

——Smith and Gasser

- ✓ **Be Multimodal**
- ✓ **Be Incremental**
- ✓ **Be Physical**
- ✓ **Explore**
- ✓ **Be Social**
- ✓ **Learn a Language**

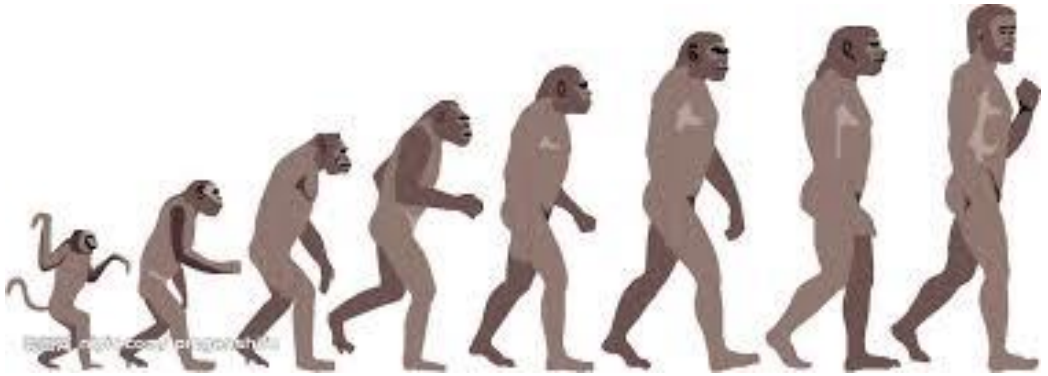
- L. Smith and M. Gasser, “The development of embodied cognition: six lessons from babies.,” Artificial life, vol. 11, no. 1-2, 2005.

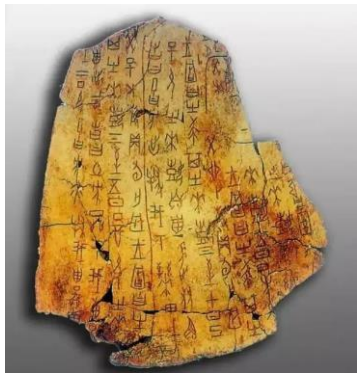
3 感知→行为：视觉语言导航

➢ 语言的重要性

语言和文字是人类思维的载体，是人类特有的智能。

- 动物有知觉、有性格、有情感、有意识
- 动物唯独没有文字





3 感知→行为：视觉语言导航

➢ 语言的重要性

语言和文字中的不确定性，无论是随机性，还是模糊性，正是语言和文字的魅力所在。

- 语音和文字的不确定
- 句义的不确定
- 词义的不确定
- 切分的不确定

苏东坡的词没有景德镇的瓷好

中国队大败（胜）美国队

你不理财，财不理你

下雨天留客天留人不留

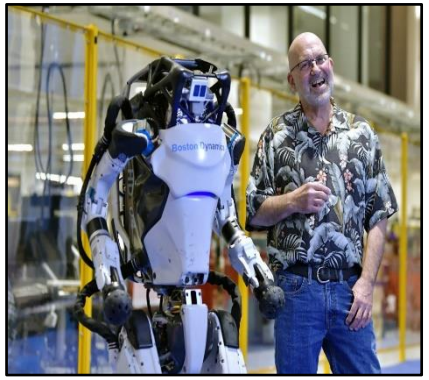
- **下雨，天留客，天留人不留！**
- **下雨天，留客，天留，人不留！**
- **下雨天，留客天，留人不？留！**
- **下雨天，留客天，留人？不留？**

3 感知→行为：视觉语言导航

➢ 语言的重要性

医护机器人

送物机器人



人形机器人

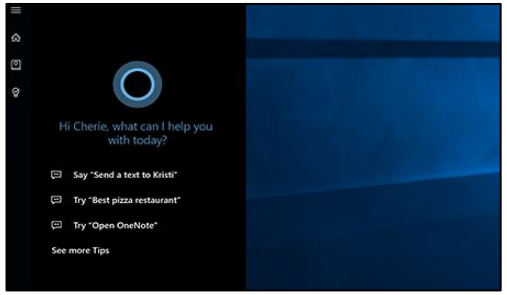
虽然能够帮助我们完成一些具体的任务，但是无法跟我们进行语言上的交流，我们需要达到一个目标，即给机器人下达一般的口头指令，并让其执行所要求的的任务。

Apple Siri

嘿 Siri, 帮我设置一个下午三点的闹钟。

百度 小度

你好小度，我要听郭德纲的相声。



Microsoft Cortana

小娜, 请问今天下午会下雨吗?

3 感知→行为：视觉语言导航

➢ 视觉语言导航

3 感知→行为：视觉语言导航

大家叫外卖的时候，会不会有时候外卖小哥明明已经到附近，但又找不到路送货，要电话联系半天，对著手机比手画脚好几分钟，才让他找到正确的道路。 外卖小哥不像我们每天走，对路途不熟悉也是应该的，但每次叫外卖都要解说一下很麻烦，甚至还要走出去接应对方。

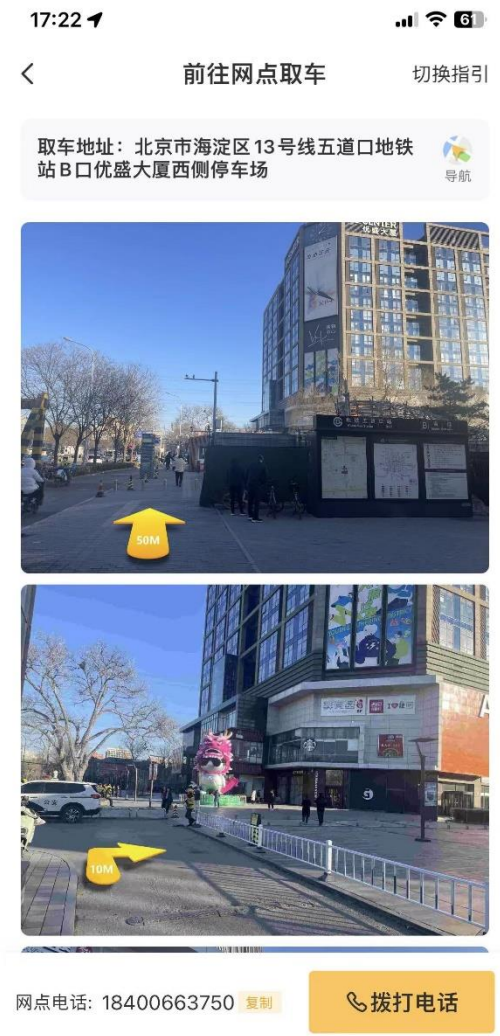
后来我发现了这方法，用了以后就可以大幅减少外卖小哥来电次数： 在收货地址最后加上送货详情：

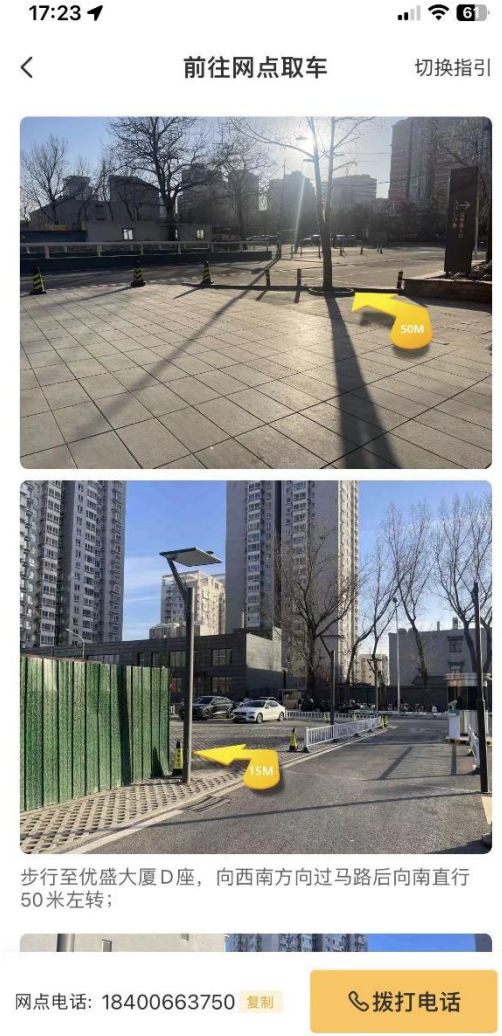
比方说： 「……G406(面向超市左侧干洗店旁电梯上四楼走廊走到底)」

「……馆(汉庭酒店XXX店坐电梯到三楼大堂通过前台右转上楼梯直上顶楼)」

根据实测，系统的格子够用，除了方便我们，也是帮外卖小哥省事。

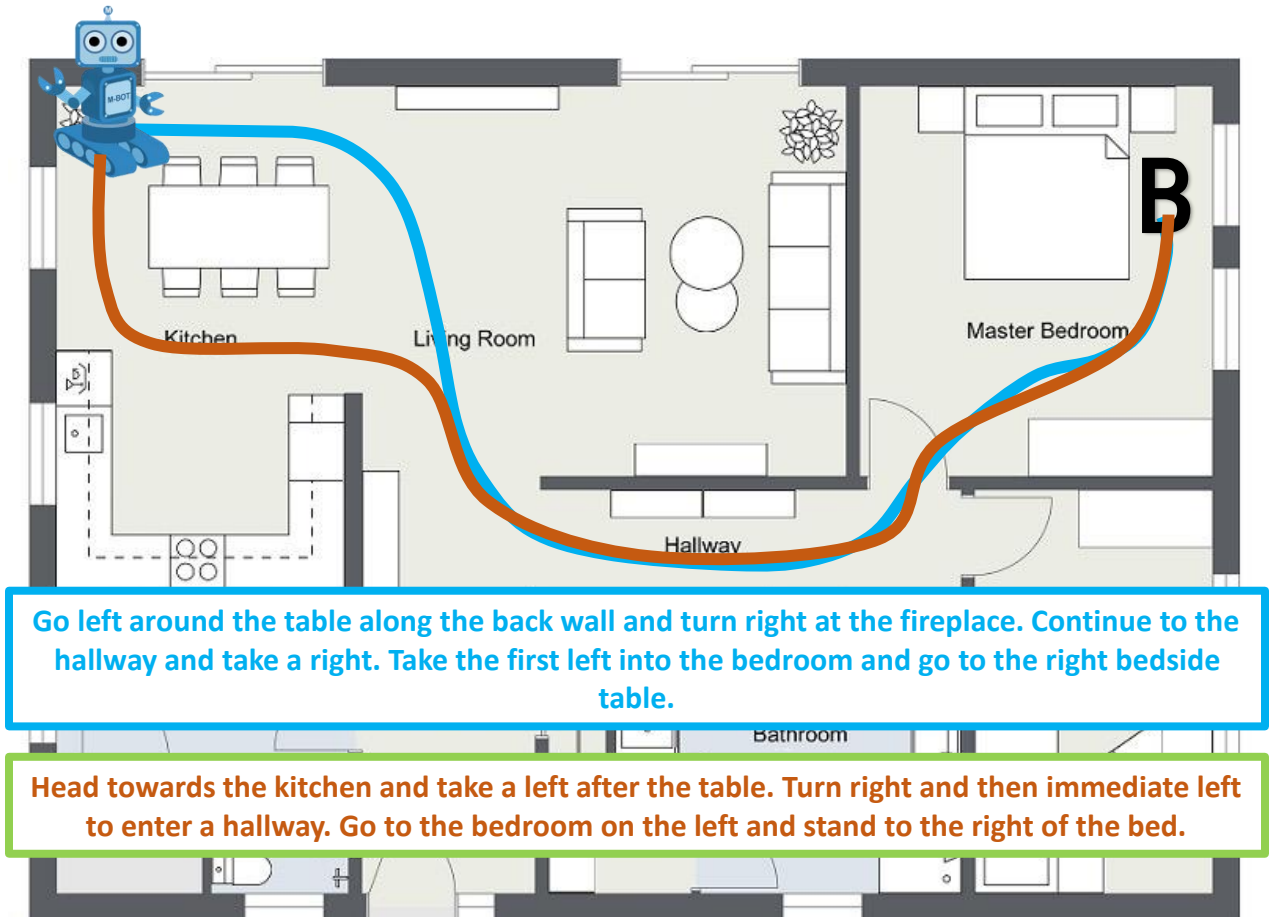
3 感知→行为：视觉语言导航





3 感知→行为：视觉语言导航

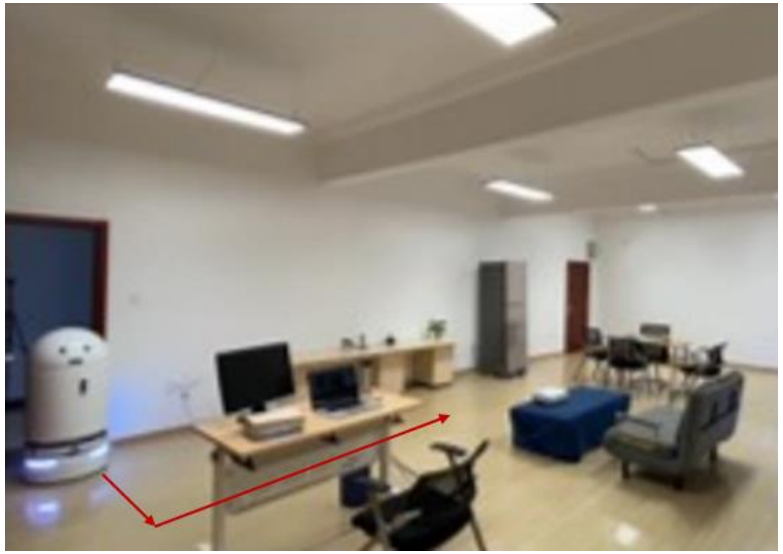
➢ 视觉语言导航



- Vision language navigation (VLN) 让智能体跟着自然语言指令进行导航，这个任务需要同时理解自然语言指令与视角中可以看见的图像信息，然后在环境中对自身所处状态做出对应的动作，最终达到目标位置。

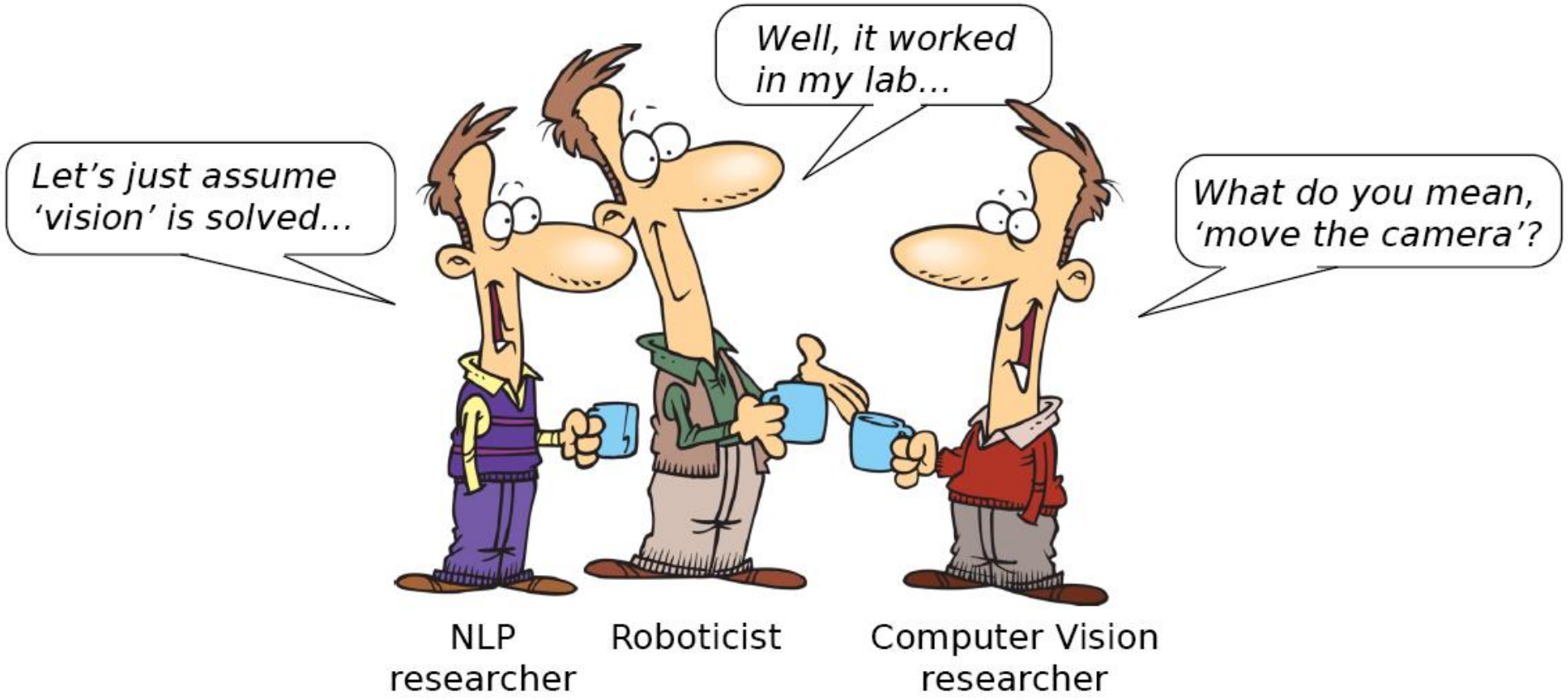
3 感知→行为：视觉语言导航

➢ 视觉语言导航



3 感知→行为：视觉语言导航

➢ 视觉语言导航：挑战



3 感知→行为：视觉语言导航

➢ 早期研究：Follow Navigational Directions

1. go vertically down until you're underneath eh diamond mine
2. then eh go right until you're
3. you're between springbok and highest view-point

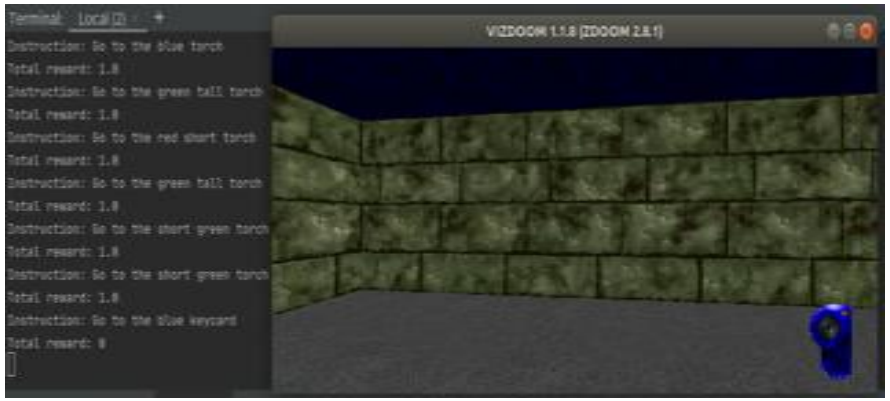
指令发送者根据面前的地图，描述为导航指令，发给指令接收者。

指令发送者和指令接收面对面，看不见彼此的地图。

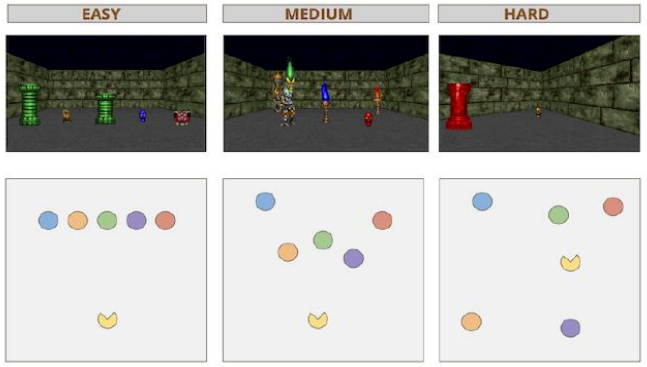
- Vogel, et al. Learning to Follow Navigational Directions. ACL2010

3 感知→行为：视觉语言导航

➢ 早期研究：Follow Navigational Directions



一个以自然语言为导向的任务，在3D毁灭战士的环境中使用语言指令。测试集由未知的指令组成。自然语言的组成是物体名称，颜色，大小信息的排列组合。



任务分为三个不同的难度级别，

- 简单:目标在智能体的视场一条水平线的五个固定位置生成。
- 中等:目标在随机位置产生，但环境确保它们在智能体的视野范围内。
- 困难:目标和智能体在随机位置生成，目标在初始配置中可能在智能体的视野中，也可能不在。智能体需要探索以寻找物体。

- Chaplot, et al. Gated-Attention Architectures for Task-Oriented Language Grounding. AAAI 2018

3 感知→行为：视觉语言导航

➢ 问题描述

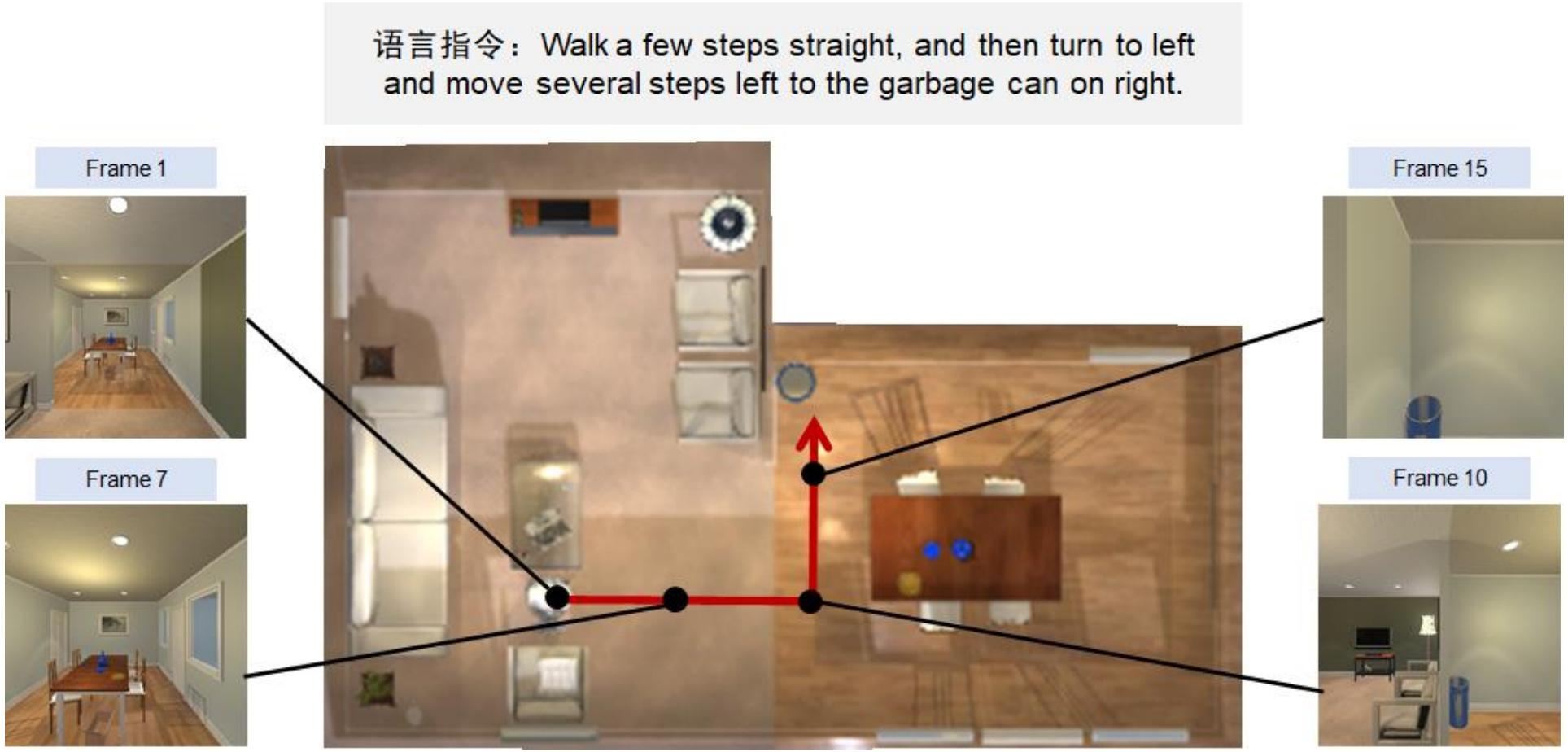
Input：

- First-person view visual Scene (no GPS, no map)
- Language instruction:

1. Walk beside the outside doors and behind the chairs across the room.
2. Turn right and walk up the stairs.
3. Stop on the seventh step.

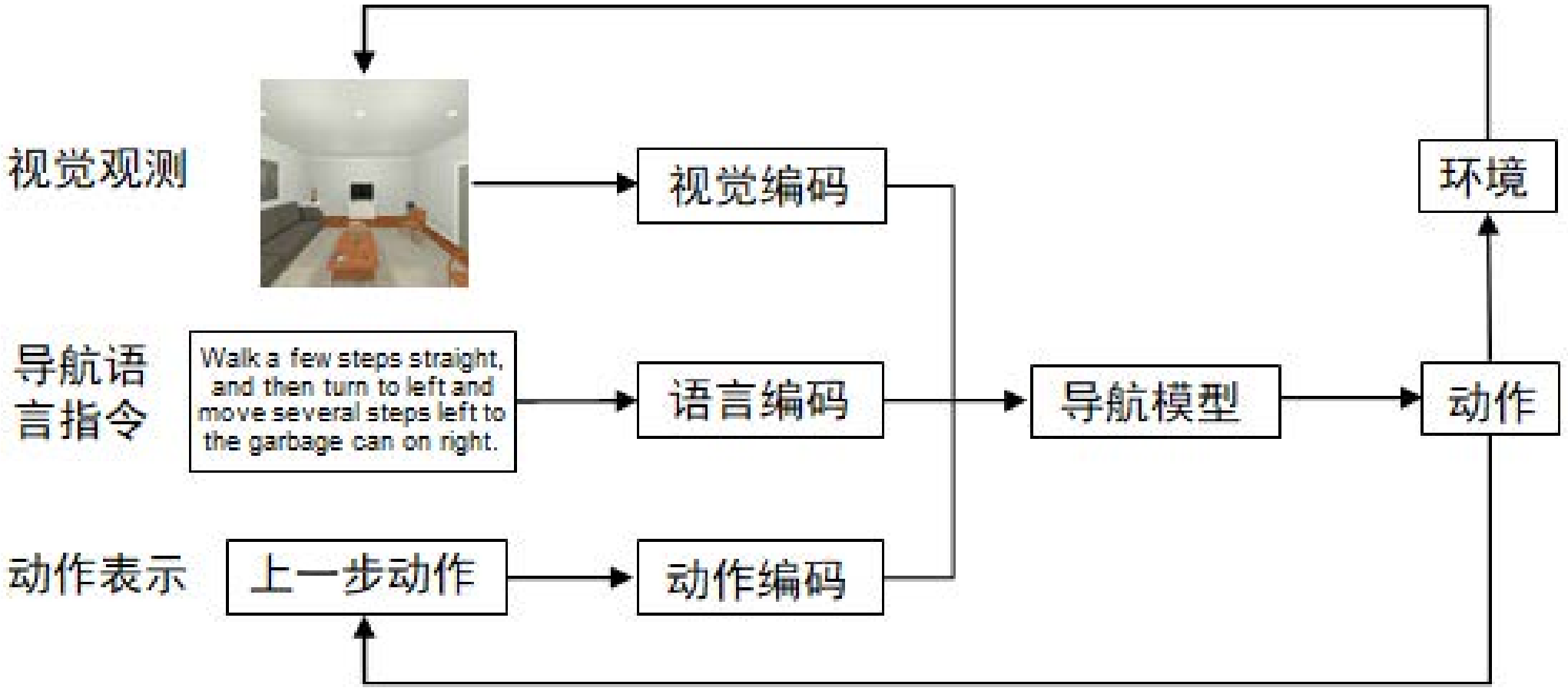
Output:

Action: Turn left, Turn right, Look up down, Go forward, Stop ...



3 感知→行为：视觉语言导航

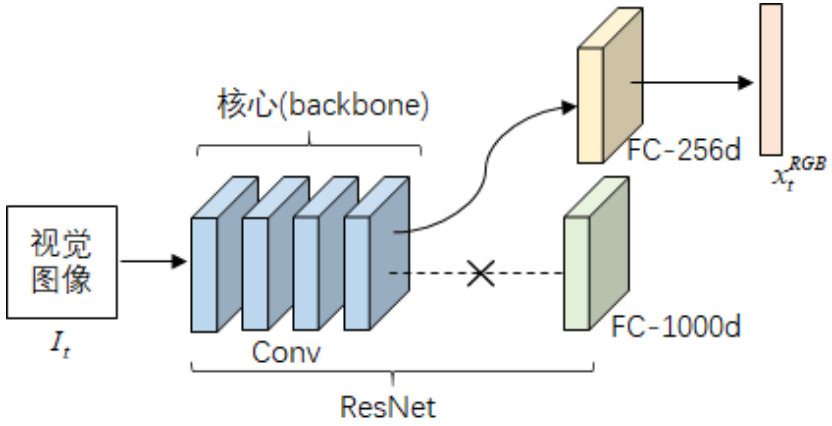
- 方法

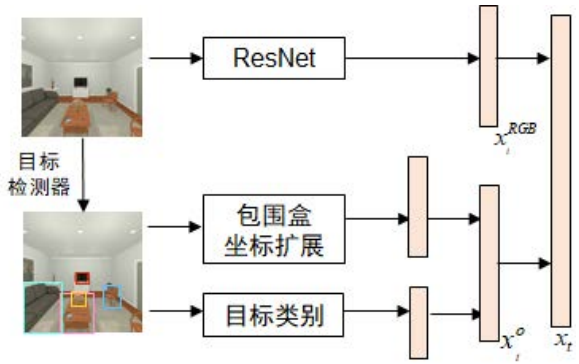


3 感知→行为：视觉语言导航

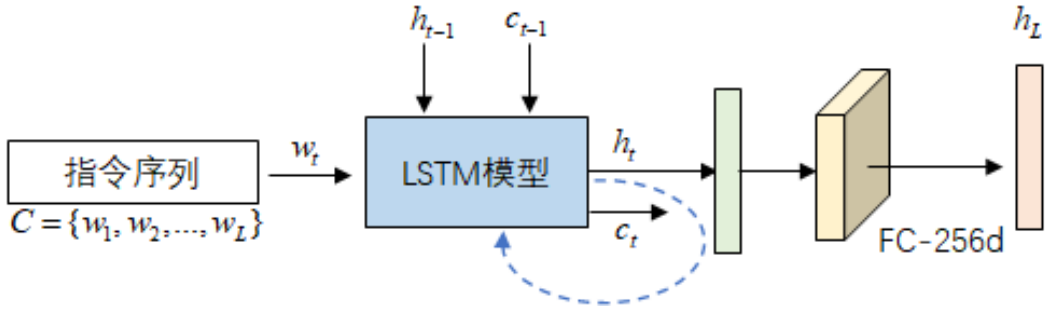
➢ 方法

视觉特征提取

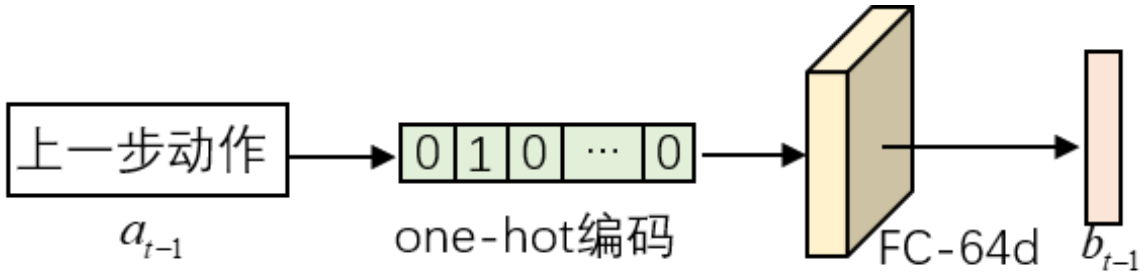




语言特征提取

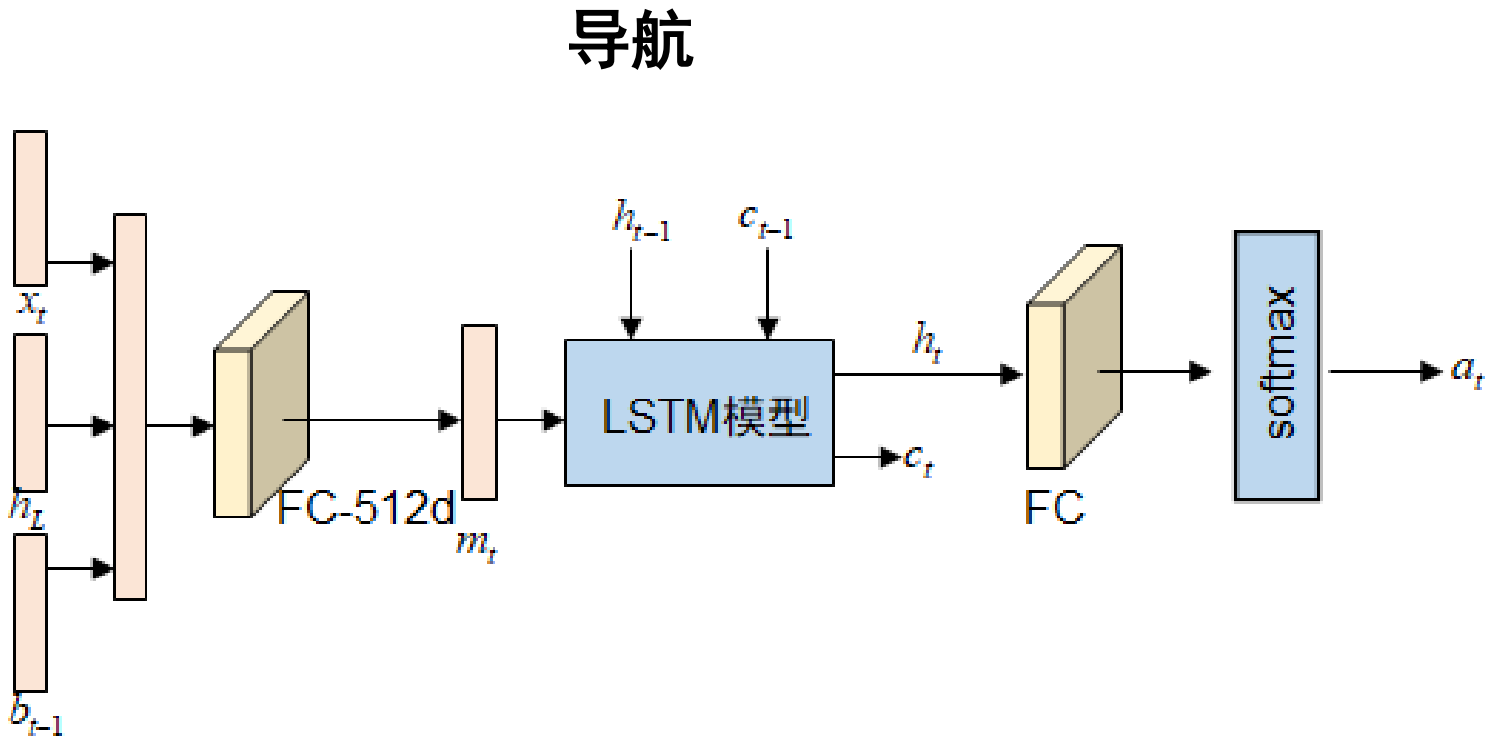


动作特征提取



3 感知→行为：视觉语言导航

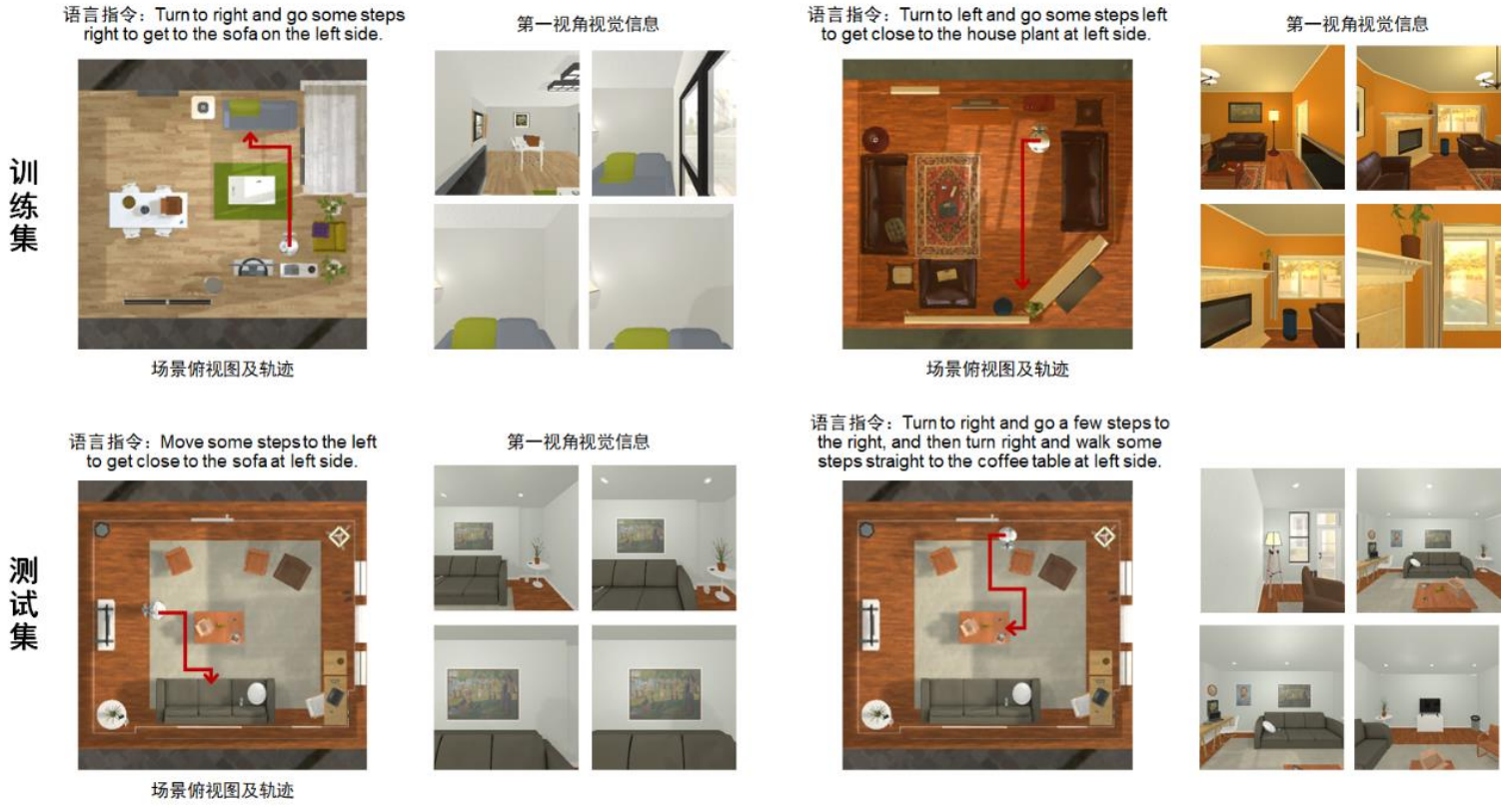
➢ 方法



$$\pi_\theta(a \mid s)$$

3 感知→行为：视觉语言导航

➢ 实例

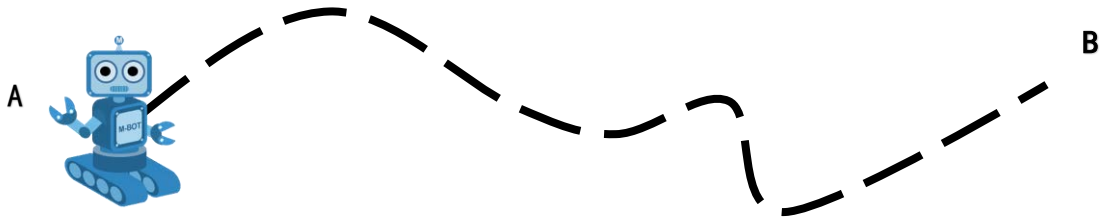


平移运动：包括停止，和向前、左、右四个方向的平移运动。每一次平移运动的移动距离固定为网格的边长 。

旋转运动：包括左转、右转两个方向的旋转运动。每一次旋转运动的旋转角度为90° 。智能体通过执行动作空间中的动作 ，不断对环境进行探索，当机器人停止动作a_stop时，结束探索

3 感知→行为：视觉语言导航

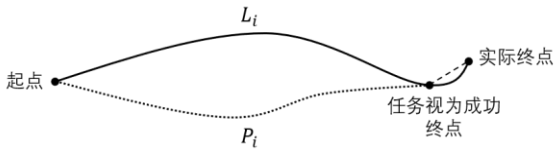
➢ 评价

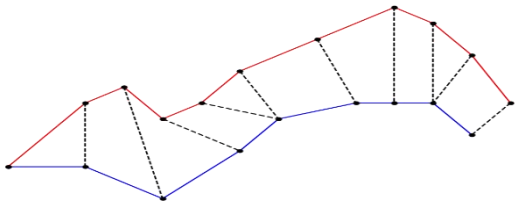


- 随机方法：按照示范路径中各动作的概率分布进行采样来决定每一步的导航动作
- IL：利用模仿学习训练动作生成模型。
- IL+Obj：引入了物体编码，采用IL的方法训练模型
- IL+SF：先通过模仿学习对模型进行训练，之后进行微调
- IL+SF+Obj：引入了物体编码，采用IL+SF的方法训练模型。

$$SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)}$$

$$nDTW(R,Q) = \exp\left(-\frac{DTW(R,Q)}{|R| \cdot d_{th}}\right)$$





	TL	SR	SPL	NE	NDTW	SDTW
Random	18.20	0.428	0.389	2.368	0.037	0.011
IL	19.50	0.541	0.445	1.925	0.205	0.139
IL+Obj	17.76	0.638	0.546	1.541	0.274	0.203
IL+SF	20.54	0.554	0.402	1.846	0.107	0.045
IL+SF+Obj	15.99	0.865	0.734	1.154	0.327	0.212

3 感知→行为：视觉语言导航

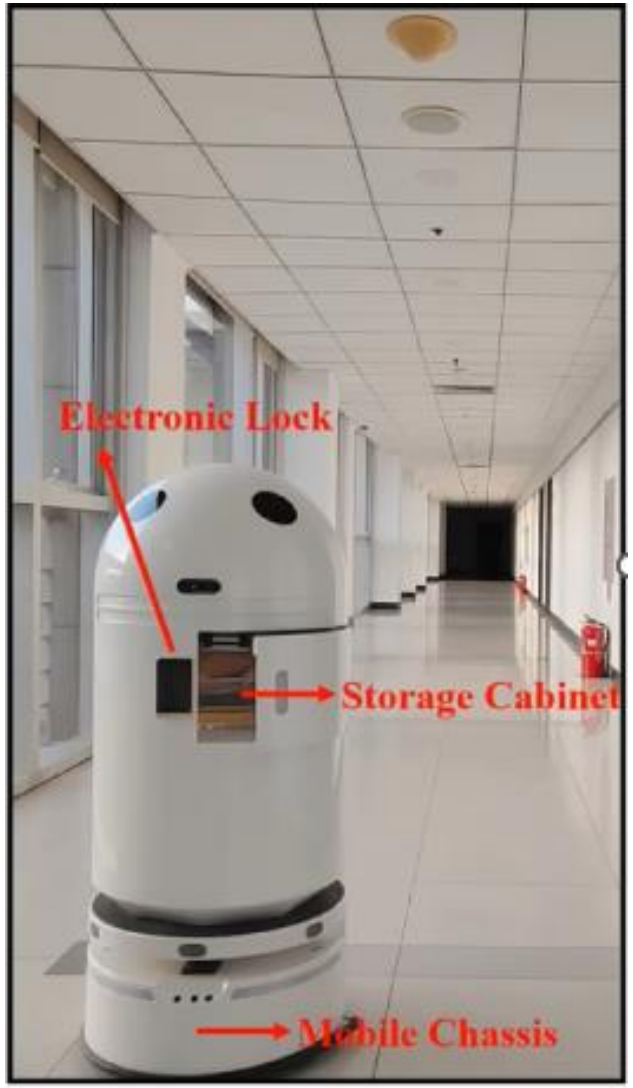
➢ 评价

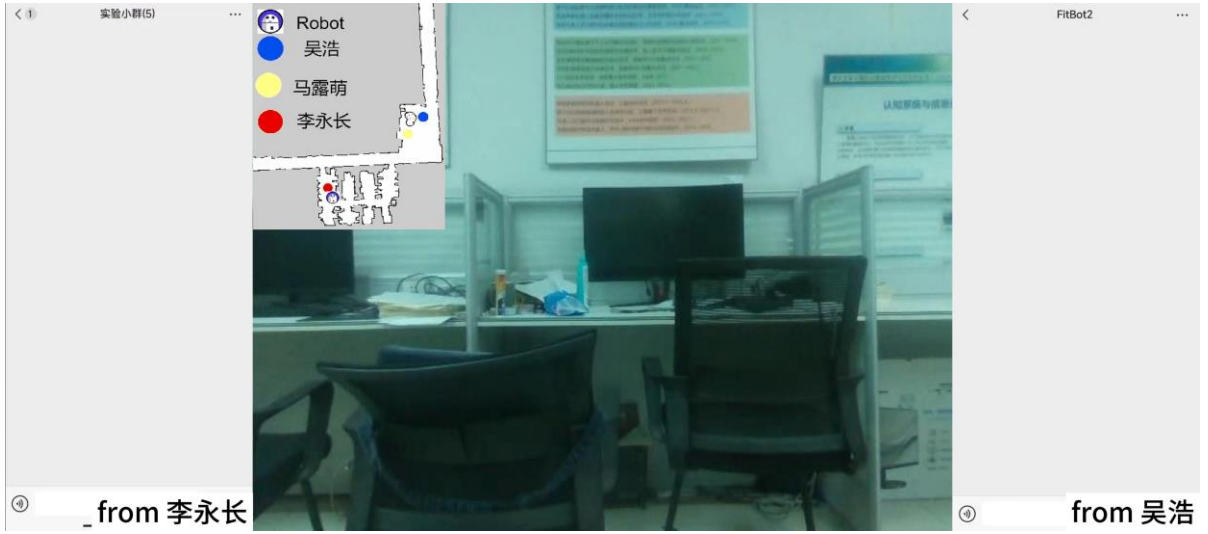
语言指令：Rotate right and go a few steps right towards the house plant on the right side.

语言指令：Turn right and move several steps to the left to get close to the light switch on right.

3 感知→行为：视觉语言导航

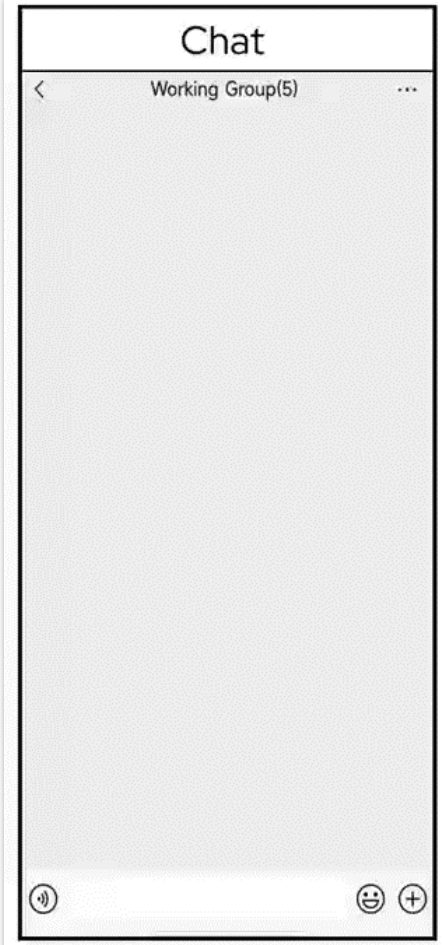
➢ 应用





3 感知→行为：视觉语言导航

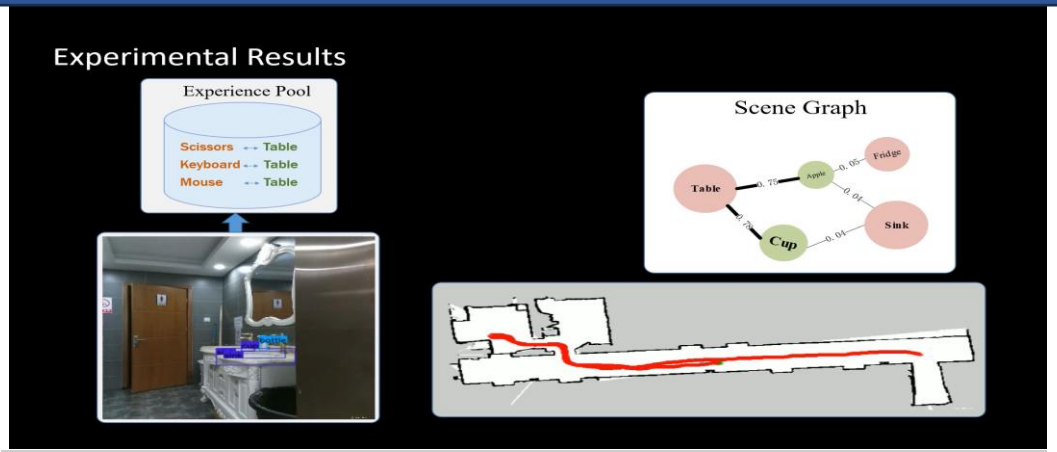
- 应用



3 感知→行为：视觉语言导航

➢ 应用：动态场景

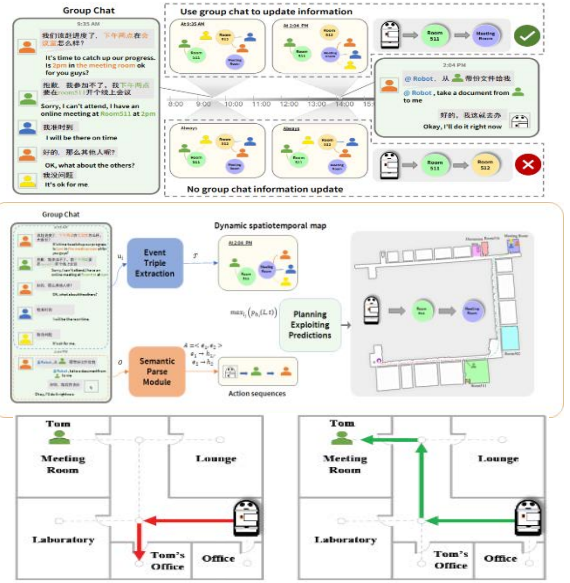
基于环境感知的动态目标晰取

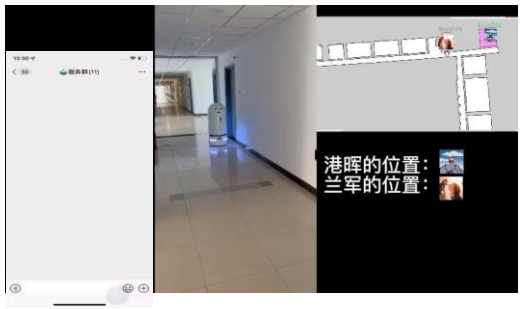


- Long-term object search using incremental scene graph updating, Robotica, 2023

动态时空场景图谱

基于聊天信息的动态目标提取





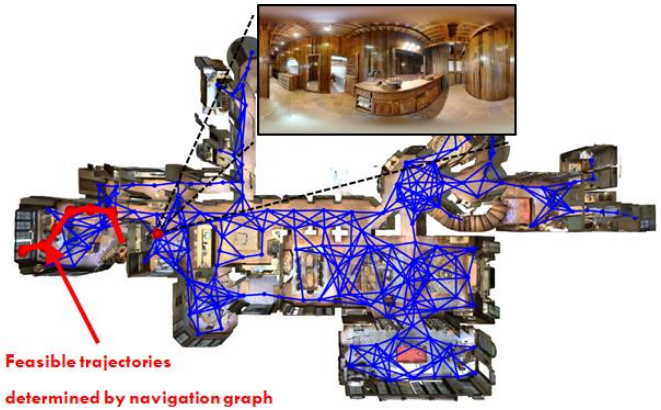
- Extracting dynamic navigation goal from natural language dialogue, IROS, 2023 (**Best Paper Award on Cognitive Robotics**)

3 感知→行为：视觉语言导航

➢ 数据集 Matterport3D Simulator for VLN Task

- Largest RGB-D dataset
- Of 90 building-scale scenes

  (avg. 23 rooms each)
- 10,800 panoramic views
- from 194,400 RGB-D images



与创建合成仿真环境相比，从真实场景中捕捉到全景图像开发的 Matterport3D 数据集能够有效地提高视觉保真度。视觉语言导航任务运行在一个固定的全景图像拓扑结构上(用蓝色表示)——假设节点之间(平均相距2.25米)有完美的导航和精确的定位。智能体执行的是高级离散的动作空间。

Chang, A, et al. Matterport3D: Learning from RGB-D data in indoor environments. 3DV 2017.

Anderson, P., Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR (2018)

3 感知→行为：视觉语言导航

➢ 数据集 Room-to-Room (R2R) Dataset

- ~7K shortest paths
- 3 instructions for each path
  - –Average instruction length 29 words
  - –Average trajectory length is 10 meters

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

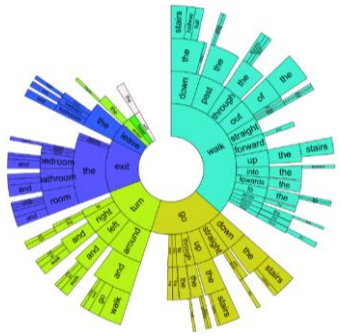
Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Input: Instruction

turn completely around until you face an open door with a window to the left and a patio to the right, walk forward, … …

Input: Panoramic View

Output



Room-to-Room (R2R) Dataset

Original Room2Room (Anderson et. al. CVPR 2018)	
Task Horizon	
Avg Words	29.4
Avg Path Len	6.0

Room4Room (Jain et. al. ACL 2019)	
Task Horizon	
Avg Words	58.4
Avg Path Len	11.1

Room6Room (Ours)	
Task Horizon	
Avg Words	**91.2**
Avg Path Len	**16.5**

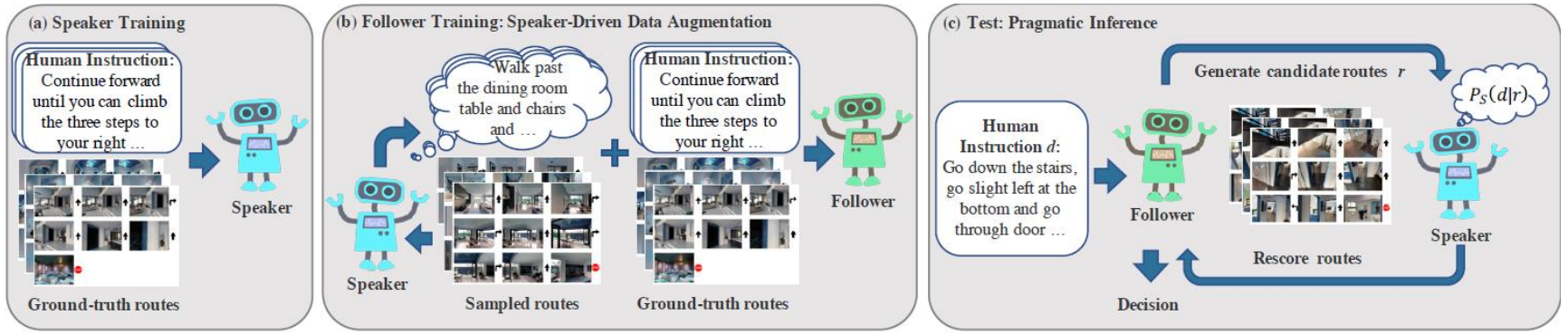
Room8Room (Ours)	
Task Horizon	
Avg Words	**121.6**
Avg Path Len	**21.6**

- Anderson, P., Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR (2018)

4 VLN 前沿

➢ Speaker-Follower

Follower：将指令映射成动作序列　　**Speaker：将动作序列映射成指令**



Speaker和Follower一起工作，共同完成导航任务，在训练时通过现有的指令和轨迹训练Speaker，合成额外的路线指令，扩展有限的训练数据，来帮助Follower。在测试阶段Follower根据不受限制的自然语言指令产生多个候选路线，由Speaker为每条路线生成对应的指令，并与正确指令比较相似性，来实际地选择最佳路线，从而提高成功的机会。

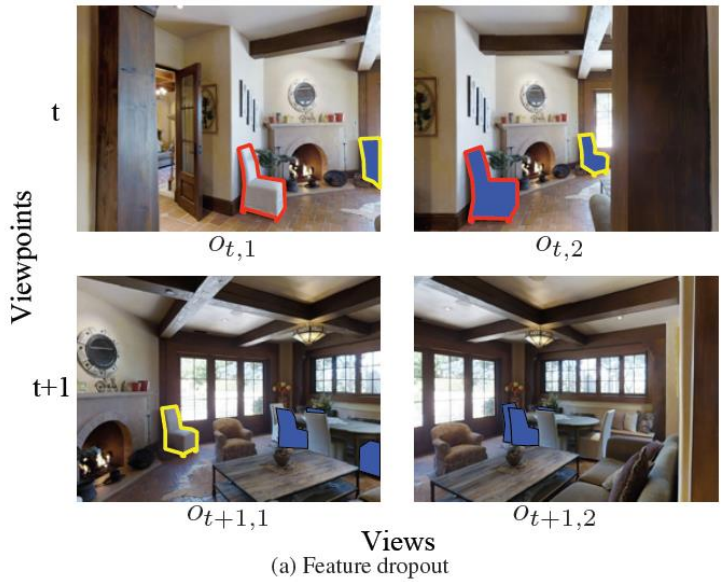
- **Fried, Daniel, et al. Speaker-Follower Models for Vision-and-Language Navigation. NIPS2018**

4 VLN 前沿

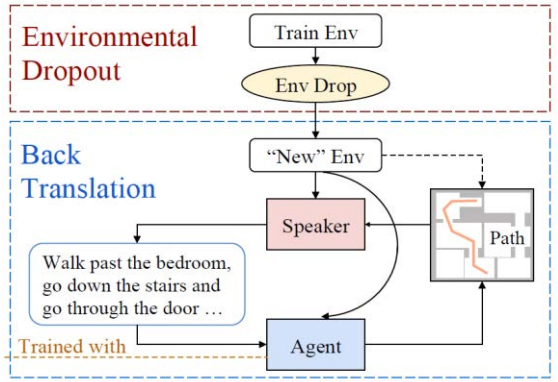
➢ Environment Dropout

从$O_{t,2}$视图中移动左椅子(标记为红色多边形);因为它也出现在$O_{t,1}$视图中，因此，Speaker仍然可以引用它，Agent也知道椅子的存在。

从O_t视图中完全移走右椅子(标记为黄色多边形)；但是在下一时刻O_{t+1}出现了这把椅子，这样的冲突信息会给Speaker和Agent造成困惑。因此不能随便地Dropout



(a) Feature dropout



Environment Dropout对环境的视觉特征做处理，随机的去移除掉该环境当中的某一类物体，以此来增加环境本身的多样性，并进一步的增加增广的数据的多样性，增强模型的泛化能力。
这两种数据增广的方法，也成了后来VLN模型的一个标配。

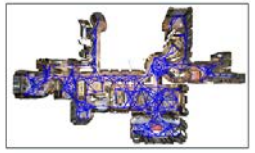
Tan Hao, et al. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. ACL2019

3.5 VLN 前沿

➢ Baby Walk

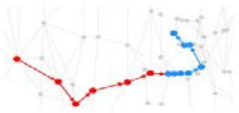
- 对数据集的长度进行扩展。采用两种方式，扩展和分割。下图是扩展，逐级增加指令和路径的长度

Original Room2Room
(Anderson et. al. CVPR 2018)



Task Horizon	
Avg Words	29.4
Avg Path Len	6.0

Room4Room
(Jain et. al. ACL 2019)



Task Horizon	
Avg Words	58.4
Avg Path Len	11.1

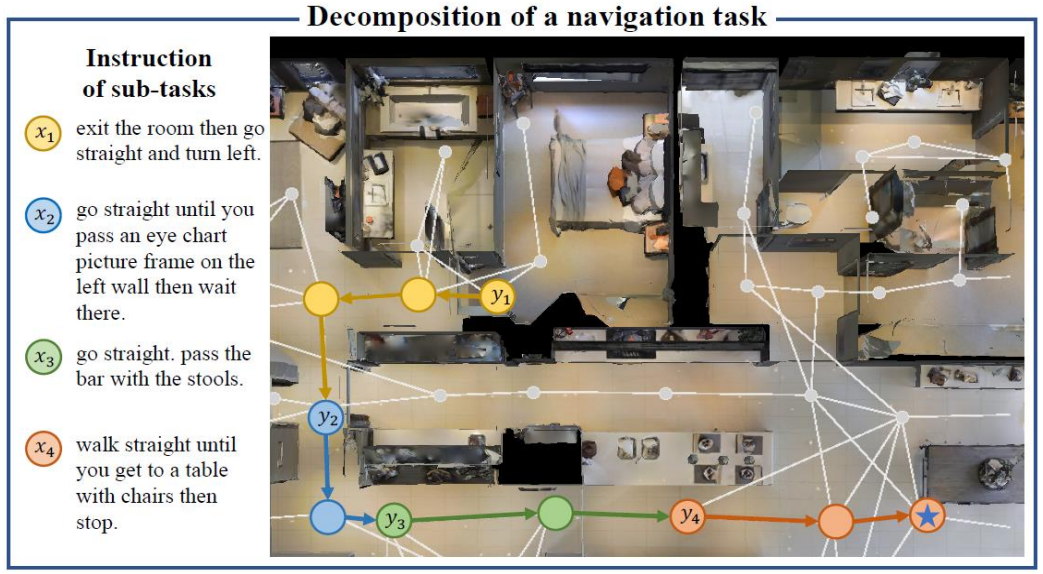
Room6Room
(Ours)

Task Horizon	
Avg Words	**91.2**
Avg Path Len	**16.5**

Room8Room
(Ours)

Task Horizon	
Avg Words	**121.6**
Avg Path Len	**21.6**

- 使用一组启发式规则从一个长指令中识别出所有可执行的婴儿步骤指令。逐步地进行训练，然后一起训练。



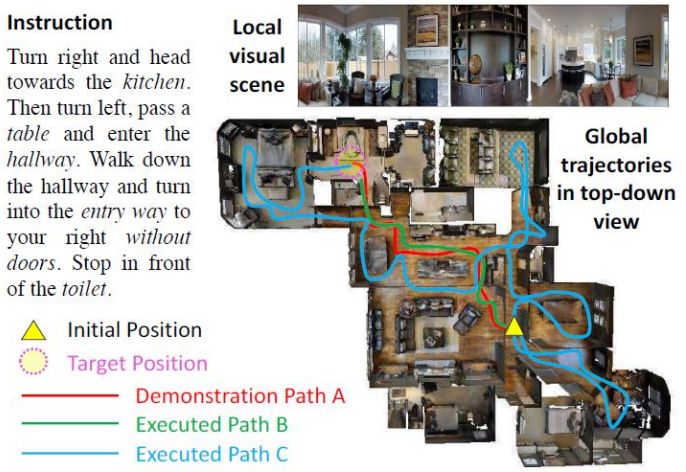
- Wang Zhu, et al. BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps. ACL 2020

3.5 VLN 前沿

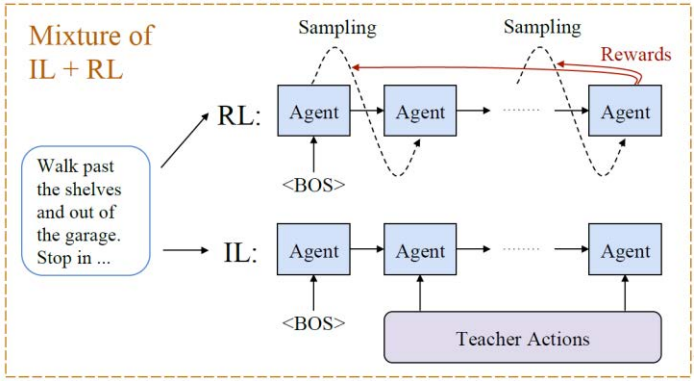
➢ RCM and SIL

VLN具有以下主要的挑战：

1. 将这些用词序列表示的指令匹配成全局时间空间中的视觉轨迹
2. 除了严格遵照专家演示之外，反馈是相当粗糙的。（对齐）
3. 泛化能力。



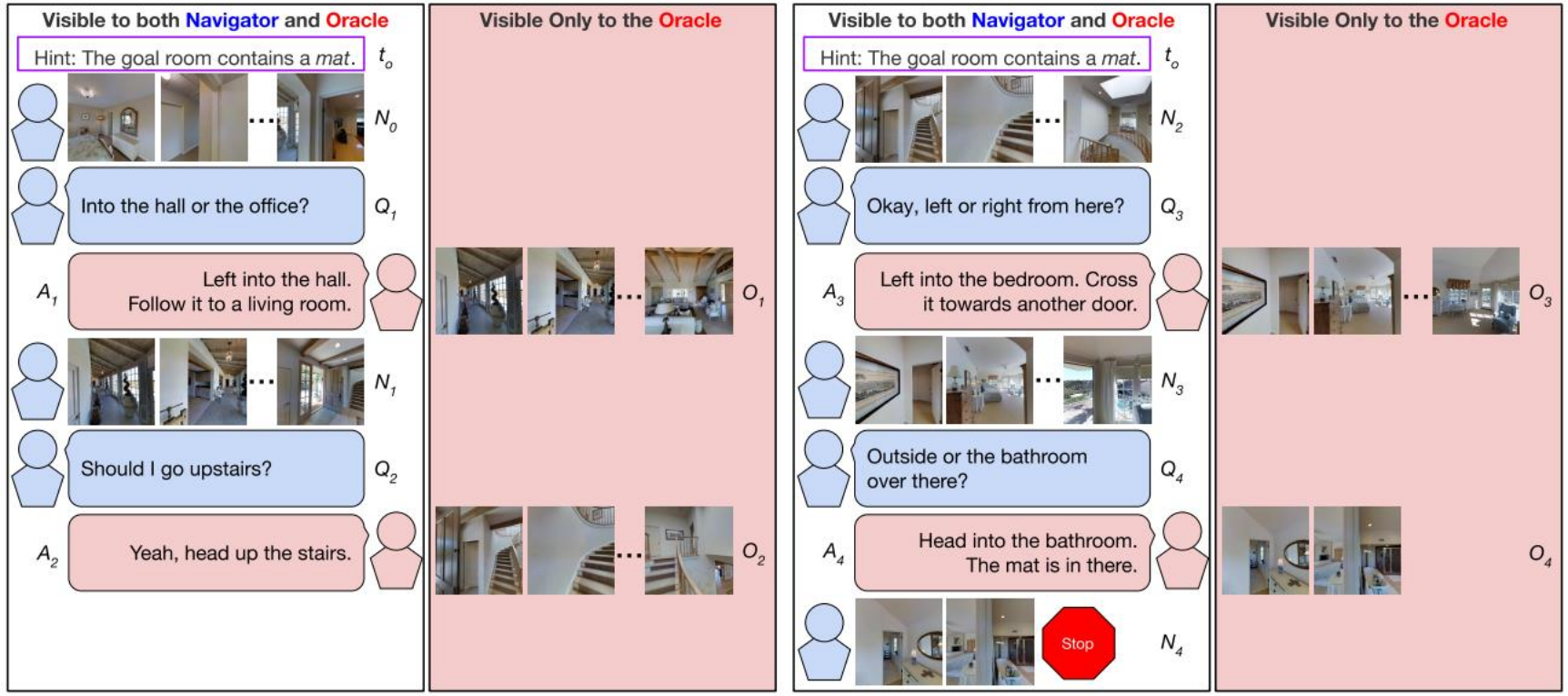
- 将Reinforcement learning (RL) 和 Imitation learning (IL) 结合在一起去训练VLN 模型，RL+IL的框架也成了后来VLN模型里的标配。
- IL依赖于将 expert 行为视为强监督信号，可以使 Agent 在已知的情况下表现更好；但是 Agent 在没有见过的环境中会遇到问题，因为每一步小错误的累积，终而酿成大错。仅使用IL的Agent在未知环境下偏向于expert行为，而不是按照指令所指示的正确路线行进。
- 结合强化学习的Agent从行为概率中抽取动作样本，并从奖励中学习，这使得智能体能够探索环境，在一定程度上提高泛化能力。



- Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. CVPR2019

3.5 前沿

➢ Dialog VLN



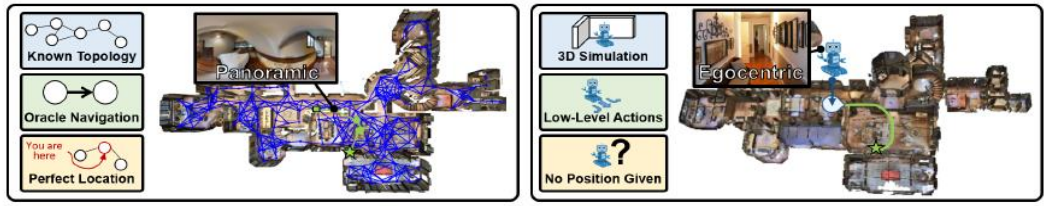
机器人在完成人类给出的导航指令时，很难一次性的根据一个复杂的指令来完成，中间可能会遇到困难，遇到ambiguous的场景和指令，那么这个时候就需要和人进行二次或多次的交互，获取更多的信息，来完成接下来的动作。

- Jesse Thomason, et al. Vision-and-Dialog Navigation, 2019, CoRL
- Nguyen, Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning，2019，EMNLP

3.5 前沿

➢ VLN-CE

- 当前Matterport3D 中进行VLN任务是根据导航图(nav-graph)一种3D空间的静态拓扑表示。导航图中的节点对应于在固定位置拍摄的360°全景图像，并且节点之间的连接表示可导航性。这种基于导航图的表述引入了许多假设，无法很好地替代智能体在现实世界中遇到的情况。



(a) Vision-and-Language Navigation (VLN) (b) VLN in Continuous Environments (VLN-CE)

Fig. 1. The VLN setting **(a)** operates on a fixed topology of panoramic images (shown in blue) – assuming perfect navigation between nodes (often meters apart) and precise localization. Our VLN-CE setting **(b)** lifts these assumptions by instantiating the task in continuous environments with low-level actions – providing a more realistic testbed for robot instruction following.

具体来讲，存在以下假设：

a) Known topology： 通过在当前全景图中选择方向并在该方向上捕捉到最近的相邻 nav-graph 节点来定义智能体的移动。实际中的智能体在新的环境中获取和更新这样的拓扑是一个开放的问题。
b) Oracle navigation： 这种节点之间的移动在感知上类似于瞬间移动，现在的全景被几米以外的新位置的全景所取代。这与一个真正的行动者在移动时所遇到的连续的观测流形成了对比。
c) Perfect localization：仿真环境中智能体随时都能得到准确的位置和方向，实际环境中，室内精确定位仍然是一个具有挑战性的问题。

Jacob Krantz, et al. Beyond the Nav-Graph: Vision-and Language Navigation in Continuous 2020, ECCV

3.5 前沿

➢ VLN-CE

然而，SLAM、路径规划等技术各自独立，远非完美，这样智能体将需要学习这些低级控制系统具有局限性，当提出的路径点不能有效到达时，将面临后果。

提出连续环境中的视觉和语言导航(VLN-CE)：

1. 智能体可以通过一系列低级动作(例如向前移动0.25米，向左转弯15度)自由地导航到任何畅通的点，而不是在固定节点之间传送。这种设置引入了以前工作中忽略的许多挑战。
2. 开发了一个简单的序列到序列基线体系结构，以及一个基于跨模态注意力的模型。
3. 使用了深度信息，经过分析表明，深度是智能体执行视觉语言导航任务不可或缺的信息。

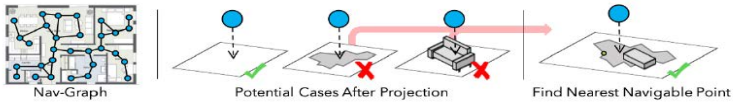
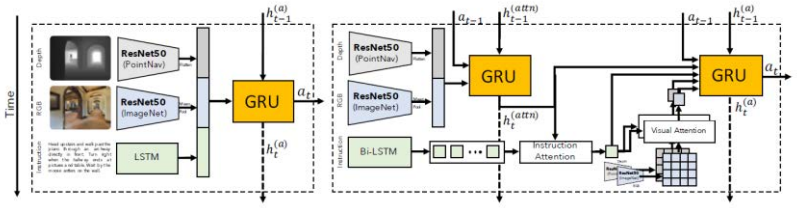


Fig. 2. We transfer nav-graph trajectories over panoramas (blue dots) from the Room-to-Room (R2R) dataset to locations in reconstructed Matterport3D (MP3D) environments. Some map to 'holes' in environment meshes where reconstruction failed or on furniture (commonly tables) where an agent could not navigate. For these, we find the nearest navigable point within 0.5m.



(a) Sequence-to-Sequence Baseline (b) Cross-Modal Attention Model

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Jacob Krantz, et al. Beyond the Nav-Graph: Vision-and Language Navigation in Continuous 2020, ECCV

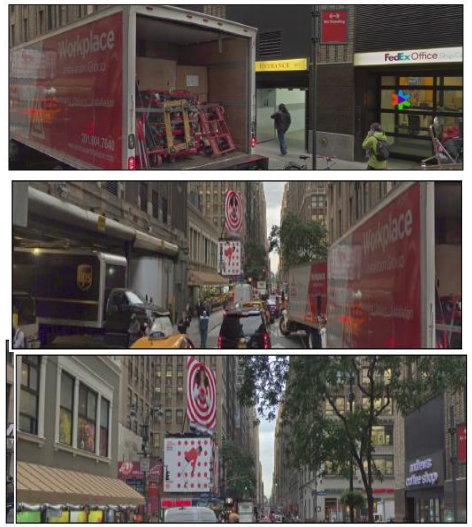
3.5 前沿

➢ Touch-Down

智能体遵循自然语言指令以实现目标，该任务分为两个部分，导航任务和空间描述方案。首先重新定位自身（顶部图像），然后继续穿过街道（两个中间图像）。在目标（底部），智能体使用空间描述（带下划线）定位“触地得分”。仅在猜测正确的情况下才会出现触地得分。

navigation tasks

Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it.

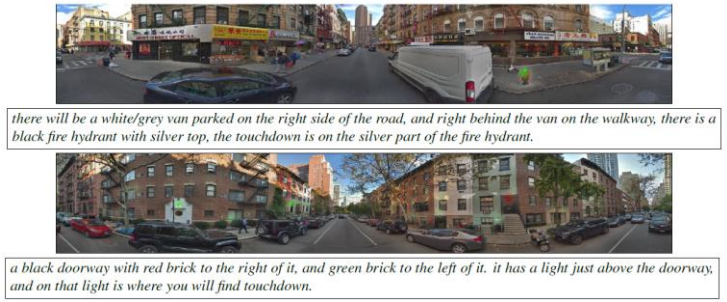


spatial description resolution (SDR)

数据集

1.模拟环境建立在Google Street View.

2.包含29641 全景图和61391条路线。



there will be a white/grey van parked on the right side of the road, and right behind the van on the walkway, there is a black fire hydrant with silver top, the touchdown is on the silver part of the fire hydrant.

a black doorway with red brick to the right of it, and green brick to the left of it. it has a light just above the doorway, and on that light is where you will find touchdown.

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

Howard Chen, et al. Touchdown Natural Language Navigation and Spatial Reasoning in Visual Street Environments, 2019,CVPR

3.5 前沿

➢ Talk2Nav

一个基于谷歌街景的交互式视觉导航环境，更重要的是设计了一种新颖的标注方法，突出了选定的地标和两者之间的空间过渡。这种增强的注释方法使众包这个复杂的注释任务成为可能。通过在 Amazon Mechanical Turk （AMT） 平台上托管这些任务，这项工作构建了一个新的数据集 Talk2Nav 。

数据集规模：包含纽约市 （NYC） 内的 10,714 远程路线。这些路由描述由 34,930 节点描述和 27,944 局部方向描述组成。每个导航指令包括5个地标描述(他们在工作中只使用4个)和4个方向说明。这5个地标描述包括关于起始路节点、目的地节点和三个中间地标的描述。由于代理从起始节点启动 VLN 任务，因此他们仅使用4个地标描述和4个方向指令。起点描述可用于自动定位，留给将来的工作。子路线的导航指令、地标描述和本地方向指令的平均长度分别为 68.8 单词、8单词和 7.2单词。总的来说，他们的数据集 Talk2Nav 包含 5.240 独特的单词。图5显示了地标描述、本地方向指令和完整导航指令的长度(字数)的分布。

(a) Top view of the navigation route

(b) Street view of the route and the agent's status

3.5 前沿

➢ REVERIE

Instruction: Bring me the bottom picture that is next to the top of stairs on level one.

REVERIE也是在MP3D数据集中构建的，机器人通过在导航图（Navigation Graph）上选择相邻的节点来移动。

相较于VLN，REVERIE的指令更加抽象，也更加接近人类所给出的真实指令。指令基本只包含了目标物体所在的房间以及其与周围物体的位置关系。

机器人需要基于当前的视觉观测进行推理，并在到达目标物体附近后，结合指令信息给出目标物体的包围盒（bounding box）。

- VLN：根据指令到达目的地。
- REVERIE：根据指令，找到目标。（其环境不可交互，不需要将物体带回）

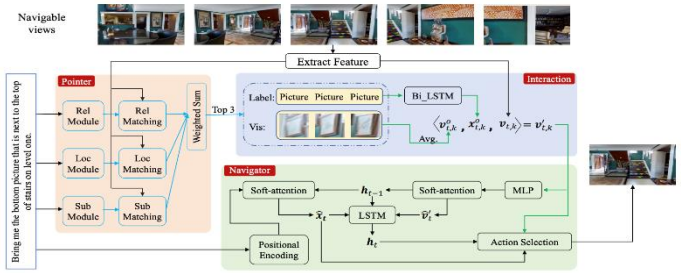


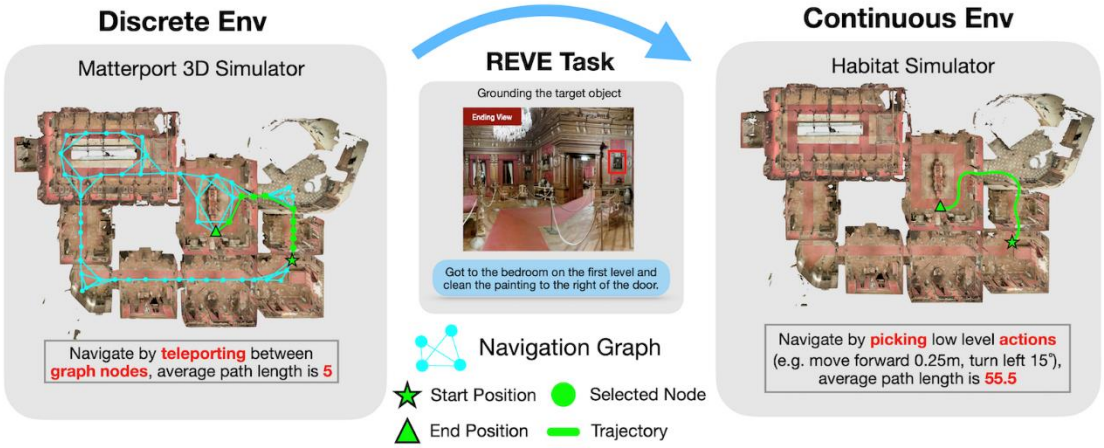
Figure 5. Our Interactive Navigator-Pointer Model

由于需要定位到指令中的目标物体，因此引入了指称表示网络MAttNet来帮助机器人对齐视觉观测中出现的物体和指令中的短语。

- REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments, 2020, CVPR

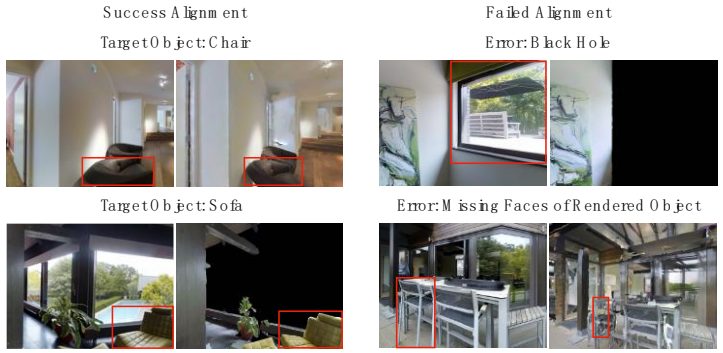
3.5 前沿

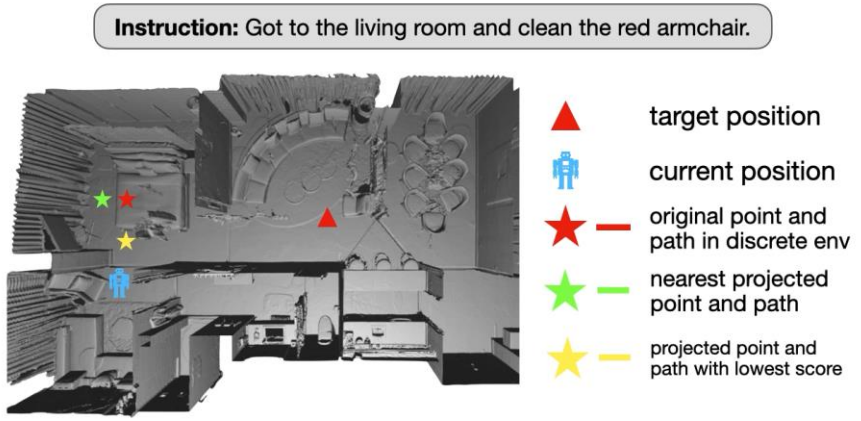
➢ REVE-CE



将REVERIE数据集从离散的MP3D环境迁移到了连续的Habitat环境中。

原始的REVERIE通过在导航图上选择相邻节点来移动，平均路径长度为5；
迁移到连续环境中的REVE-CE通过执行前进和旋转动作来移动，平均路径长度为55





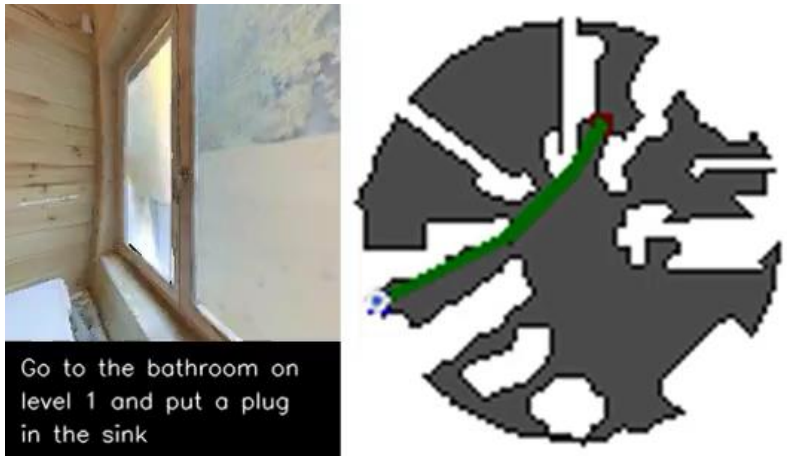
- REVE-CE: Remote Embodied Visual Referring Expression in Continuous Environment, 2022, ICRA

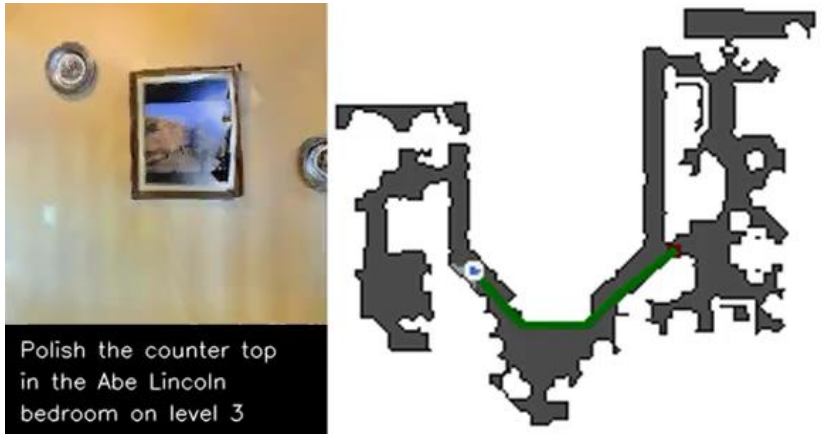
3.5 前沿

➢ REVE-CE

插值得到连续路径

基于机器人的导航动作空间，根据A*算法在相邻路点中进行搜索，即可得到完整的连续导航路径



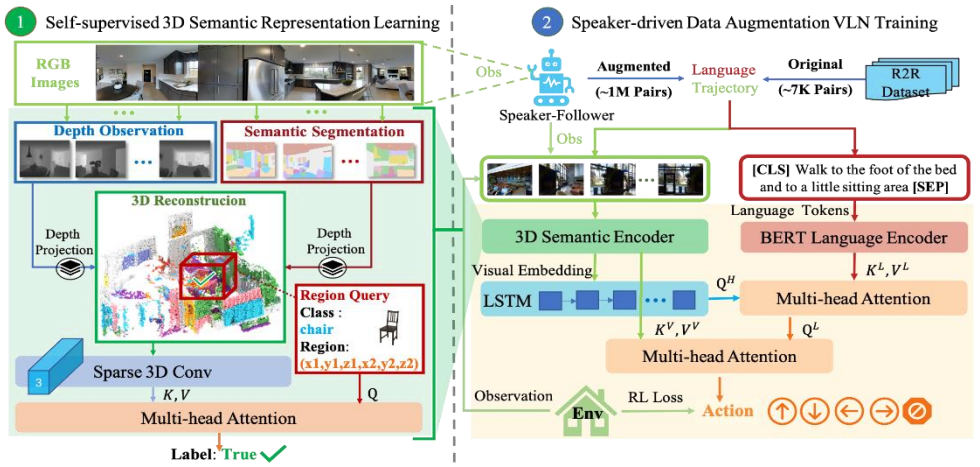


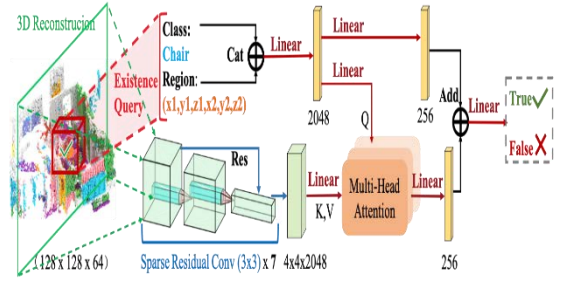
- REVE-CE: Remote Embodied Visual Referring Expression in Continuous Environment, 2022, ICRA

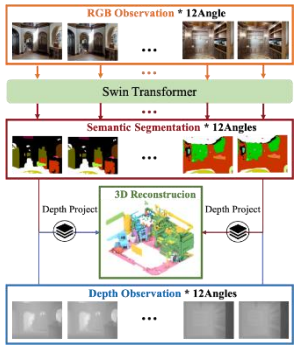
3.5 前沿

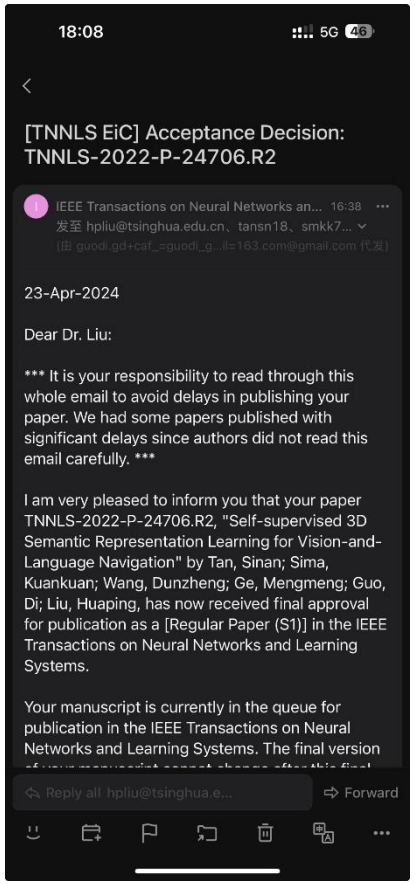
➢ 基于三维语义特征的VLN

- 将RGB-D输入对局部场景观测进行三维语义重建
  - **通过对深度信息进行反向投影，得到体素级别的三维语义重建。**
- 利用自监督学习方法转换为三维语义编码
  - **通过区域查询任务，用自监督预训练的方法得到场景的三维语义表示。**
- 导航模型：将BERT语言特征和三维语义特征特征通过导航LSTM进行融合
- 利用跨模态蒸馏对RGB特征和三维语义表示进行融合









- Self-supervised 3D semantic representation learning for vision-and-language navigation, IEEE Transactions on Neural Networks and Learning Systems, 2024

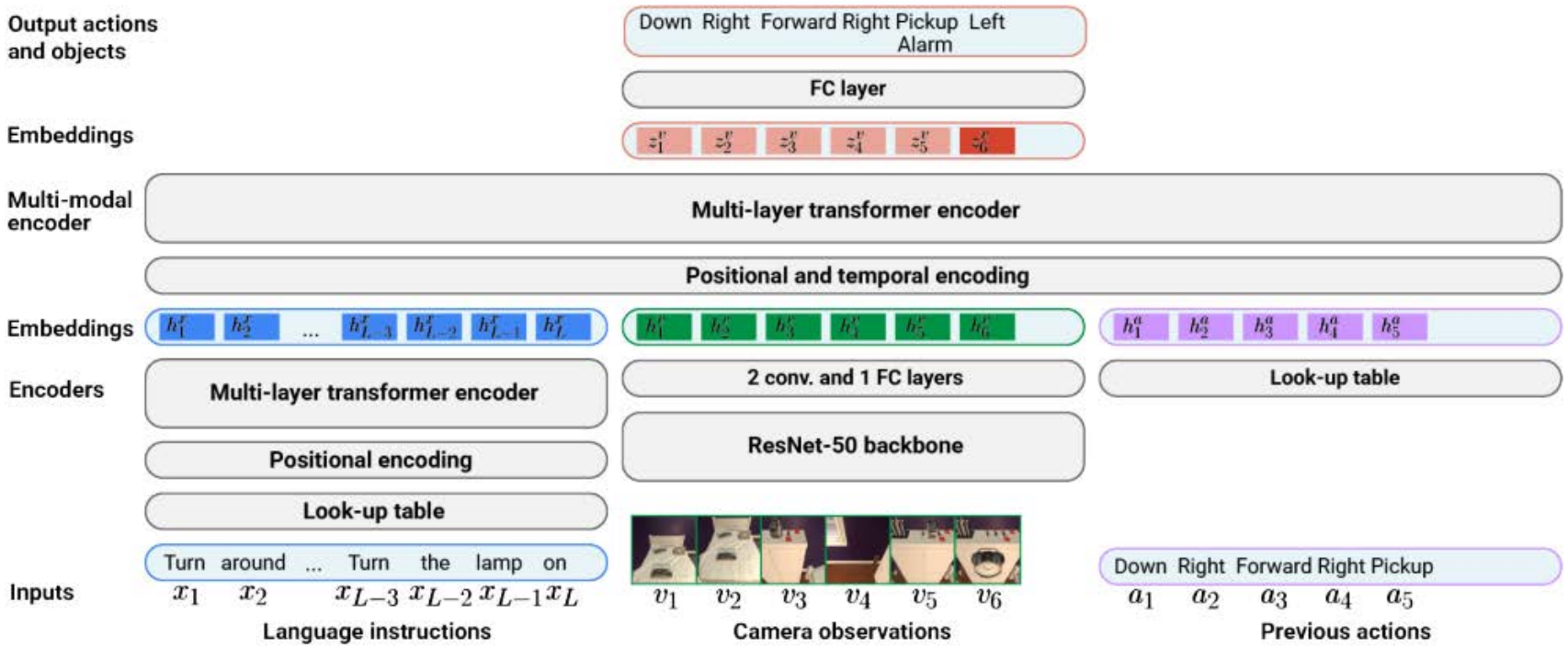
3.5 前沿

➢ 基于三维语义特征的VLN

Methods	R2R Validation Seen			R2R Validation Unseen			R2R Test Unseen		
	NE↓	SR↑	SPL↑	NE↓	SR↑	SPL↑	NE↓	SR↑	SPL↑
Random [1]	9.45	16	-	9.23	16	-	9.79	13	12
Human [1]	-	-	-	-	-	-	1.61	86	76
Methods using LSTM-based navigation models.									
Seq-to-Seq [48]	6.01	39	-	7.81	22	-	7.85	20	18
Speaker-Follower [19]	3.36	66	-	6.62	35	-	6.62	35	28
Chasing Ghosts [34]	7.59	34	30	7.20	35	31	7.83	33	30
SMNA [16]	3.22	67	58	5.52	45	32	5.67	48	35
RCM+SIL(train) [17]	3.53	67	-	6.09	43	-	6.12	43	38
Regretful Agent [49]	3.23	69	63	5.32	50	41	5.69	48	40
FAST [50]	-	-	-	4.97	56	43	5.14	54	41
Active Perception [51]	3.20	70	52	4.36	58	40	4.33	60	41
EGP [52]	-	-	-	4.83	56	44	5.34	53	42
PRESS [53]	4.39	58	55	5.28	49	45	5.49	49	45
EnvDrop [20]	3.99	62	59	5.22	52	48	5.23	51	47
SERL [54]	32	69	64	4.74	56	48	5.63	53	49
OAAM [55]	-	65	62	-	54	50	-	53	50
CMG-AAL [56]	2.74	73	69	4.18	59	51	4.61	57	50
AuxRN [18]	3.33	70	67	5.28	55	50	5.15	55	51
RelGraph [57]	3.47	67	65	4.73	57	53	4.75	55	52
NvEM [58]	3.44	69	65	4.27	60	55	4.37	58	54
NvEM+SEvol [59]	3.56	67	63	3.99	62	57	4.13	62	57
Ours	**2.55**	**74**	**70**	**3.33**	**68**	**61**	**3.73**	**66**	**60**
Methods using Transformer-based navigation models.									
DASA [12]	3.76	67	64	4.77	58	54	5.11	54	52
SSM [60]	3.10	71	62	4.32	62	45	4.57	61	46
PREVALENT [23]	3.67	69	65	4.71	58	53	5.30	54	51
ORIST [24]	-	-	-	4.72	57	51	5.10	57	52
SOAT [61]	3.63	63	58	4.28	59	53	-	-	-
AirBERT [26]	2.68	75	70	4.01	62	56	4.13	62	57
Recurrent VLN BERT [25]	2.90	72	68	3.93	63	57	4.09	63	57
HOP [62]	2.72	75	70	3.80	64	57	3.83	64	59
ADAPT(ResNet-152) [63]	2.54	76	72	3.77	64	58	3.79	65	59
DUET [28]	-	-	-	3.31	**72**	60	3.65	69	59
HAMT [27]	2.51	76	72	**2.29**	66	61	3.93	65	60
EnvEdit [22]	**2.32**	77	**74**	3.24	69	**64**	**3.59**	68	**64**
DUET*	2.60	76	72	3.49	69	59	3.77	68	58
Ours + DUET	2.35	**80**	73	3.36	71	58	3.73	**70**	58

- Self-supervised 3D semantic representation learning for vision-and-language navigation, IEEE Transactions on Neural Networks and Learning Systems, 2024

➢ Episodic Transformer

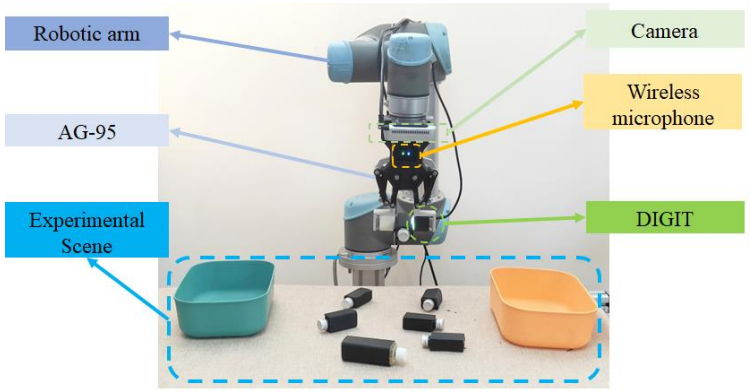


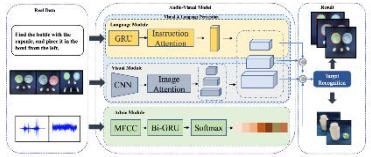
提出了一个经典的结合文字+图像+其他信息的符合Transformer模型，几乎可以作为一切VLN任务的baseline

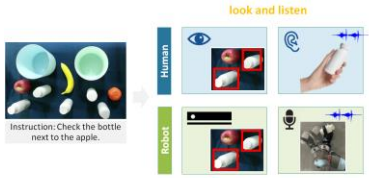
- **Episodic Transformer for Vision-and-Language Navigation，ICCV2021**

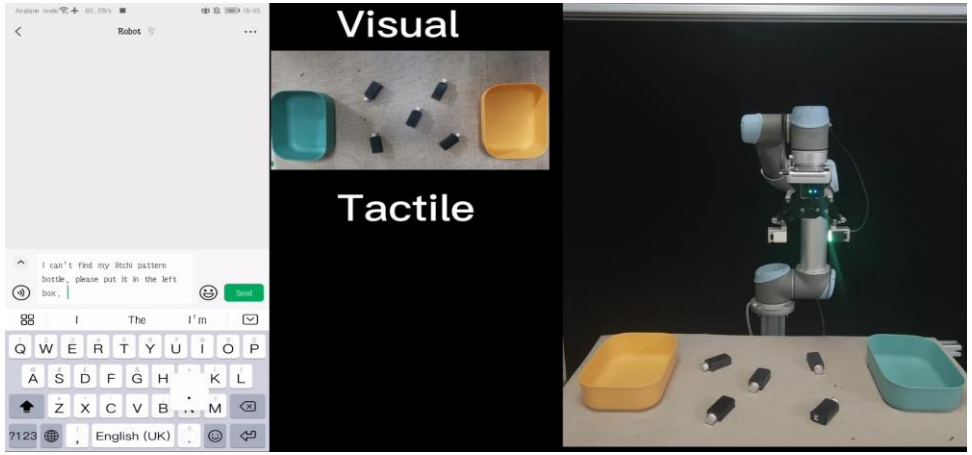
3.5 感知→行为：视觉导航

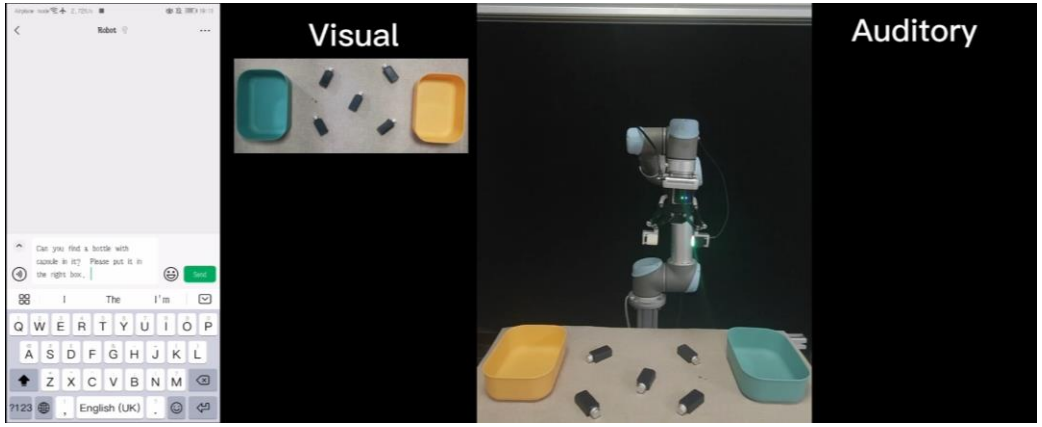
➢ 前沿：多模态（视、听、触觉与语言）引导的操作

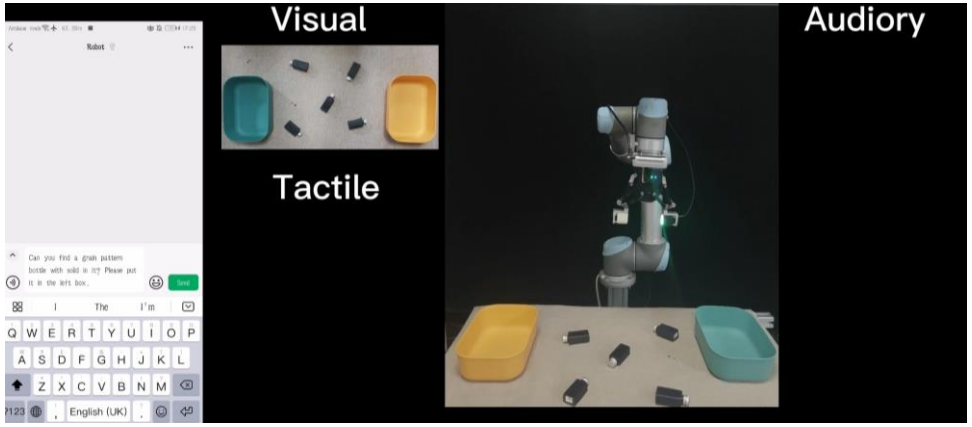








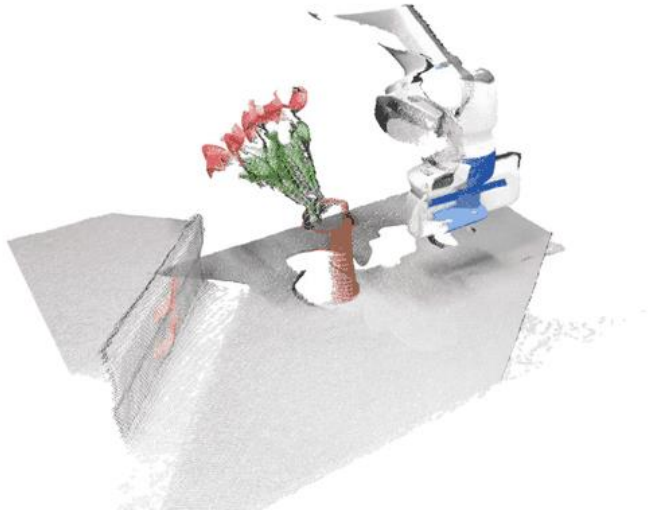


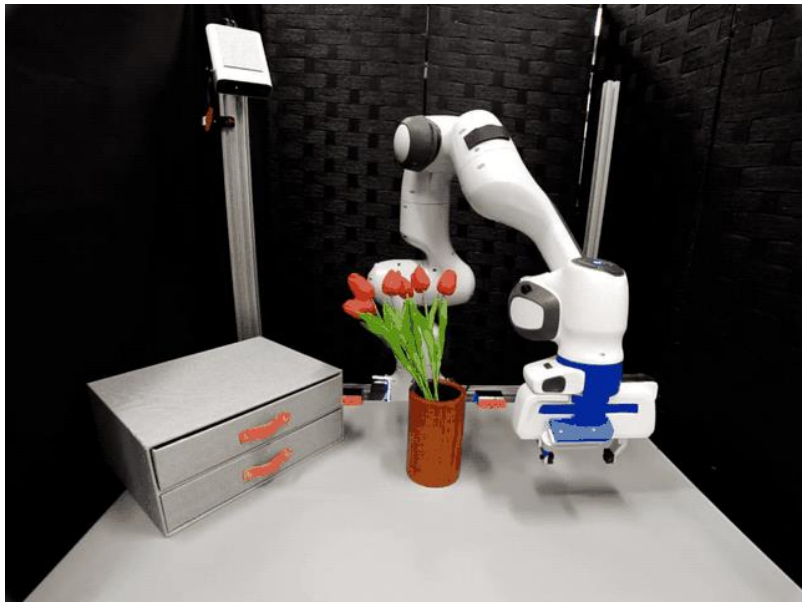


- Natural language instruction understanding for robotic manipulation: a multisensory perception approach, ICRA, 2023

3.5 前沿

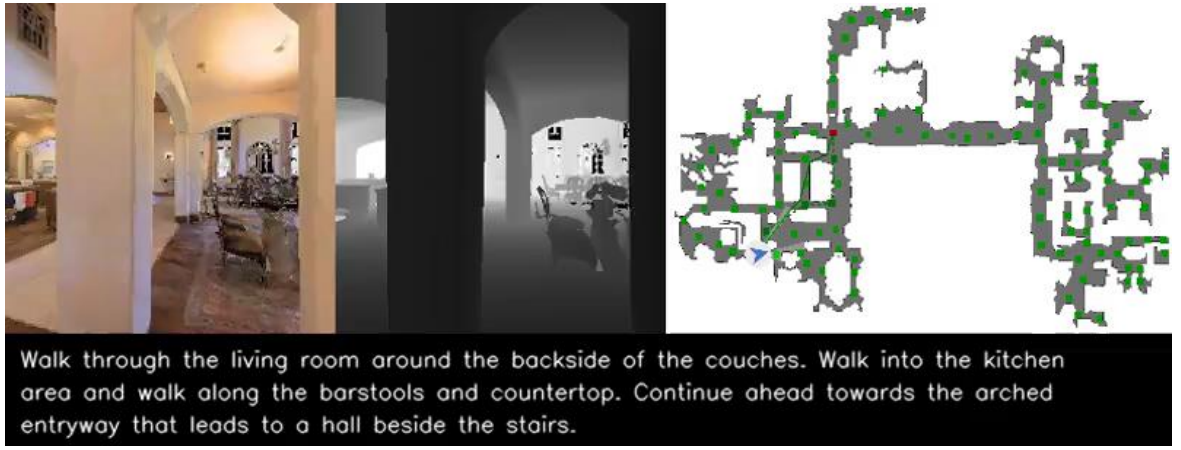
➢ 大模型—VoxPoser





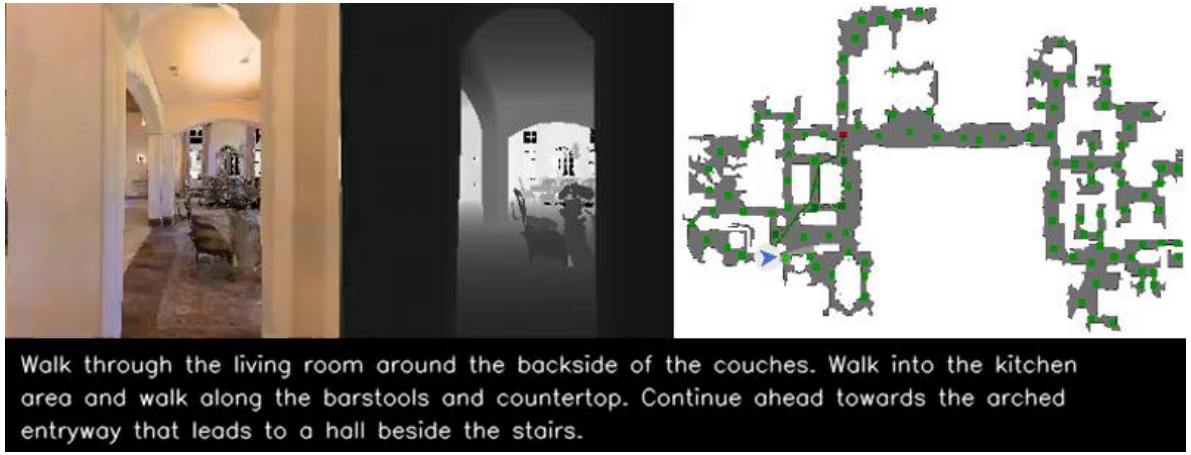
4 具身智能的脆弱性

forward_0.25m SPL = 0.82



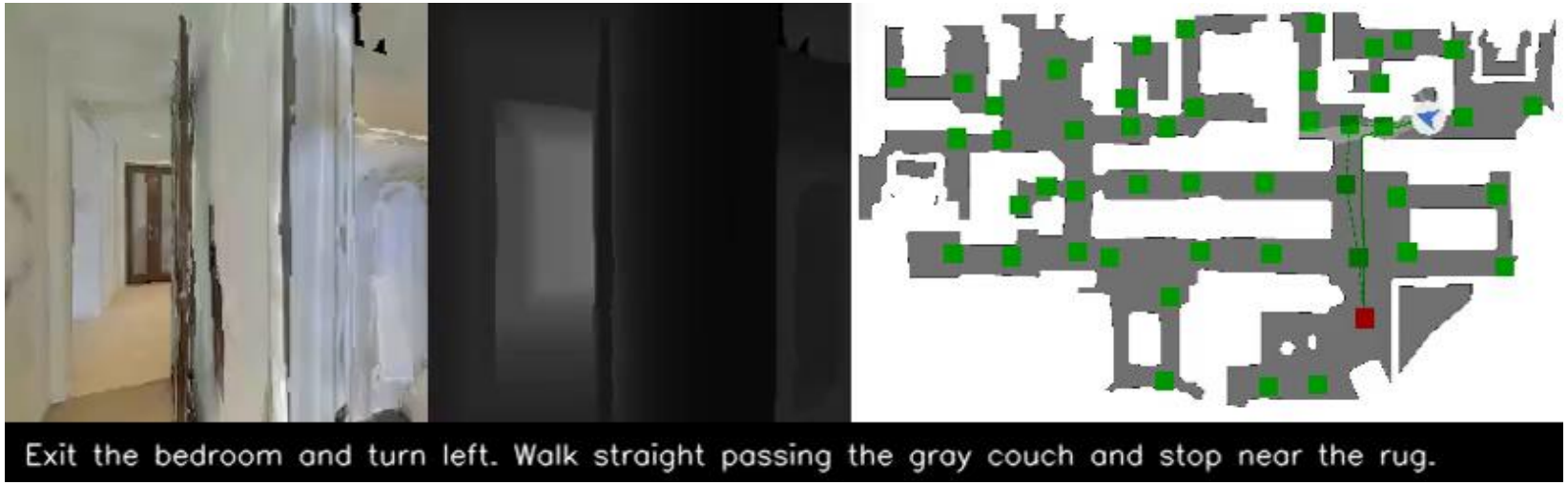
turn angle 15° SPL = 0.00

forward_0.05m SPL = 1.00



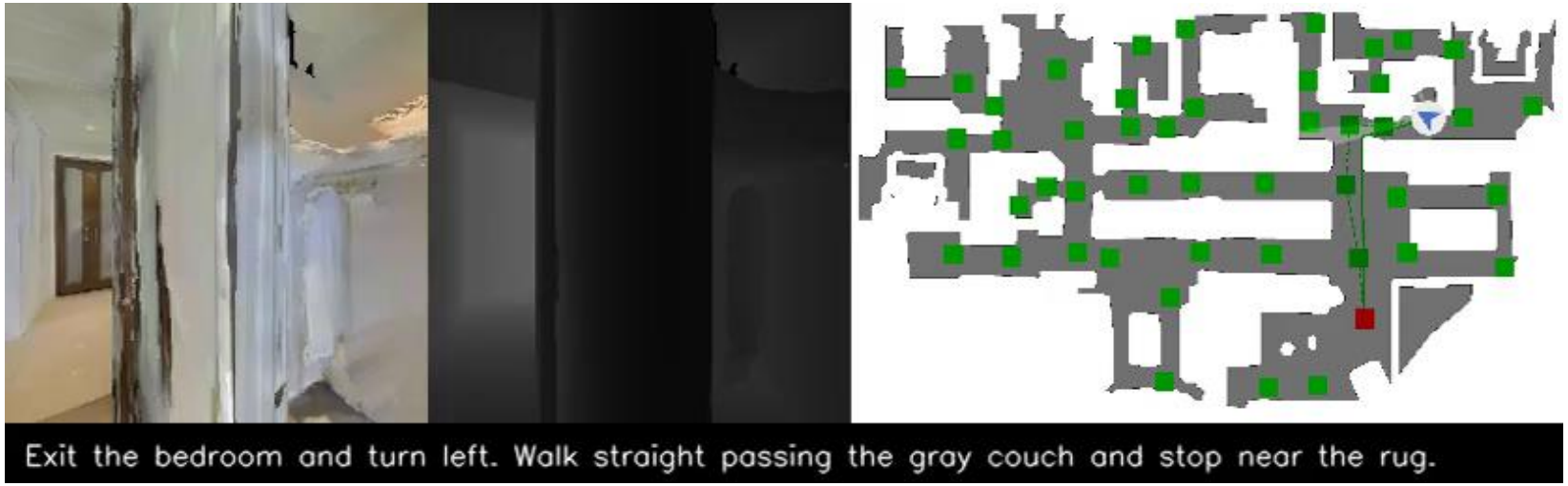
turn angle 5° SPL = 0.85

4 具身智能的脆弱性



turn angle 15°

SPL = 1.00



turn angle 1°

SPL = 0.65

4 具身智能的脆弱性

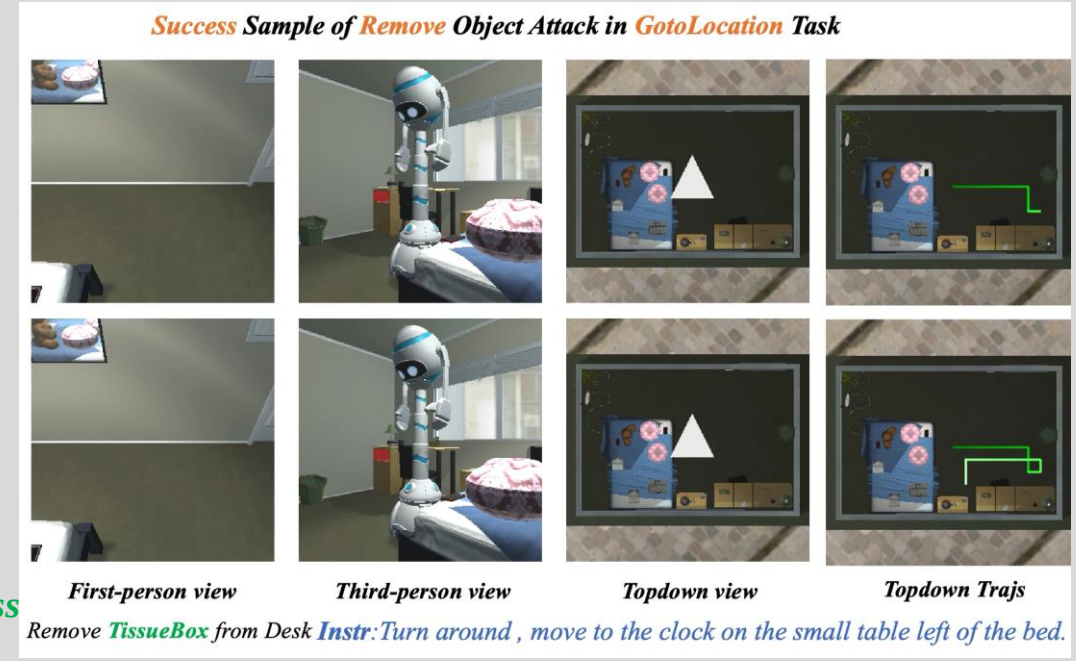
VLN

Rem

35 Steps Fail

Not Rem

40 Steps Success

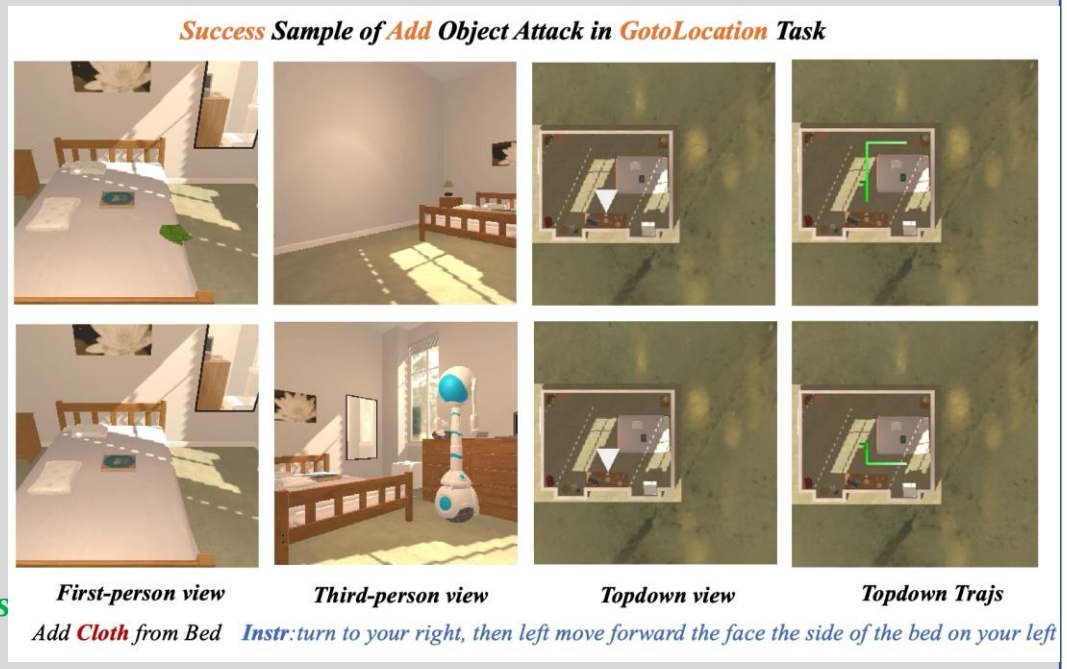


Add

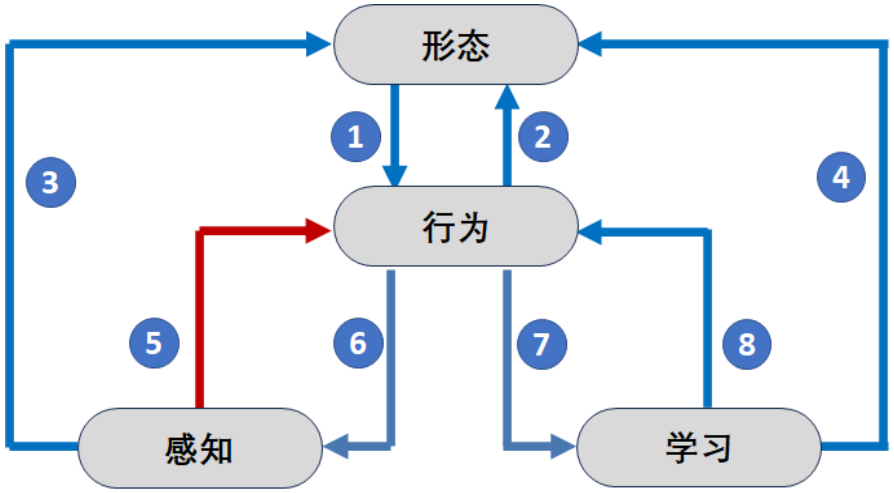
29 Steps Fail

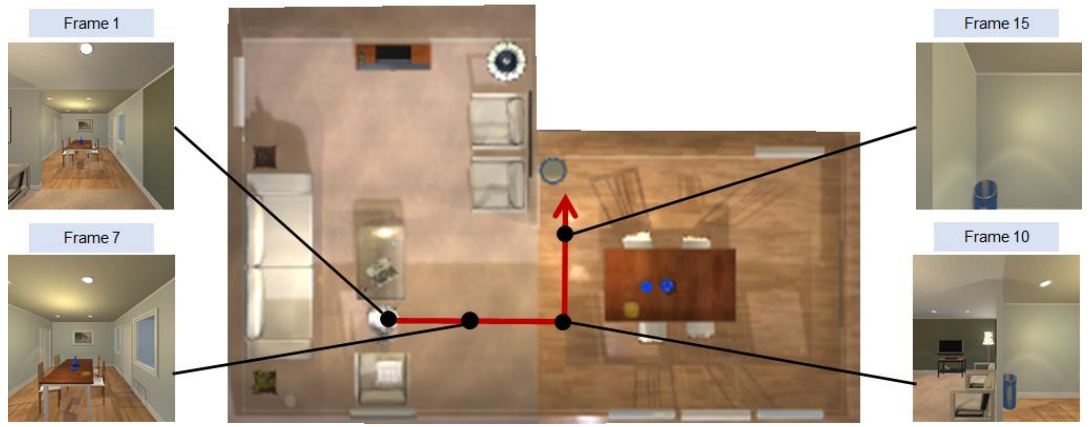
Not Add

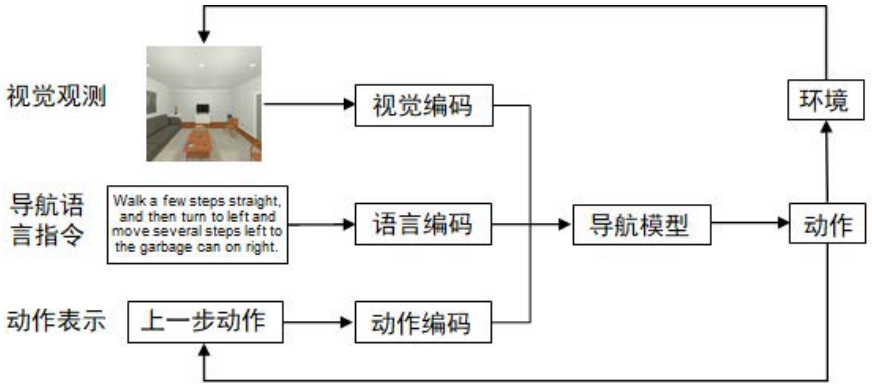
13 Steps Success

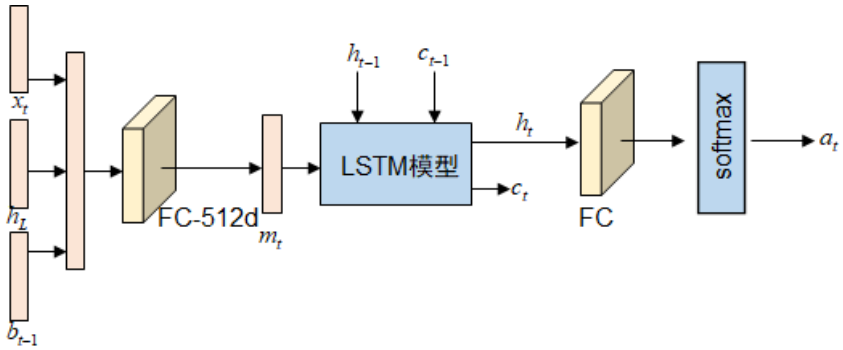


小结









- 视觉-语言-动作对齐、连续场景，… …

小结