具身智能-10

刘华平

2025年4月23日

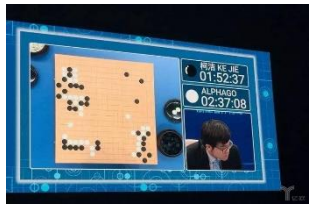
课程内容安排

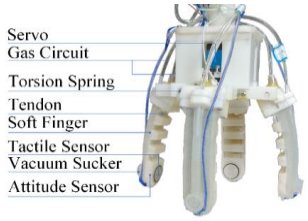
课次	周次	上课内容	软件
1	1	绪论	
2	2	深度学习	
3	3	强化学习1	Gym, Mujuco
4	4	强化学习2	Gym, Mujuco
	5	作业准备	
5	6	自监督与持续学习	
	7	开题	Powerpoint
6	8	形态智能	Gym, Mujuco
7	9	视觉导航：VLN	AI2THOR
8	10	主动感知：VSN，EQA	AI2THOR
	11	五一放假	
9	12	具身学习	AI2THOR
10	13	多体智能	AI2THOR
11	14	面向具身智能的AIGC	AI2THOR
	15-16	成果准备与展示	Powerpoint

具身智能的体系

- 狭义与广义的具身智能

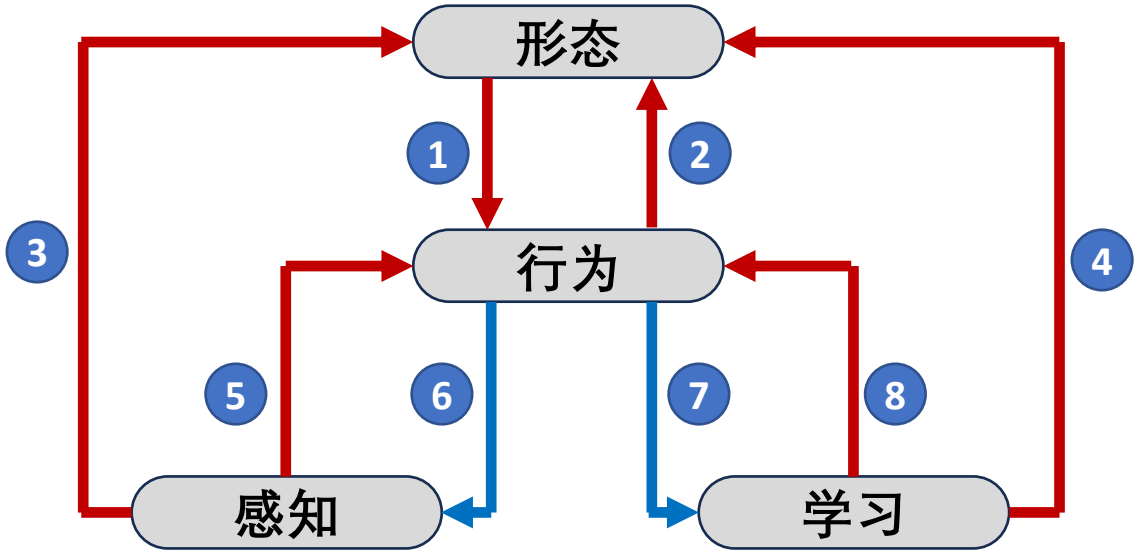
离 身 智 能 具 身 智 能

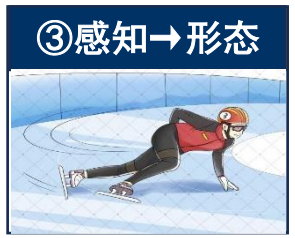




具身智能的体系

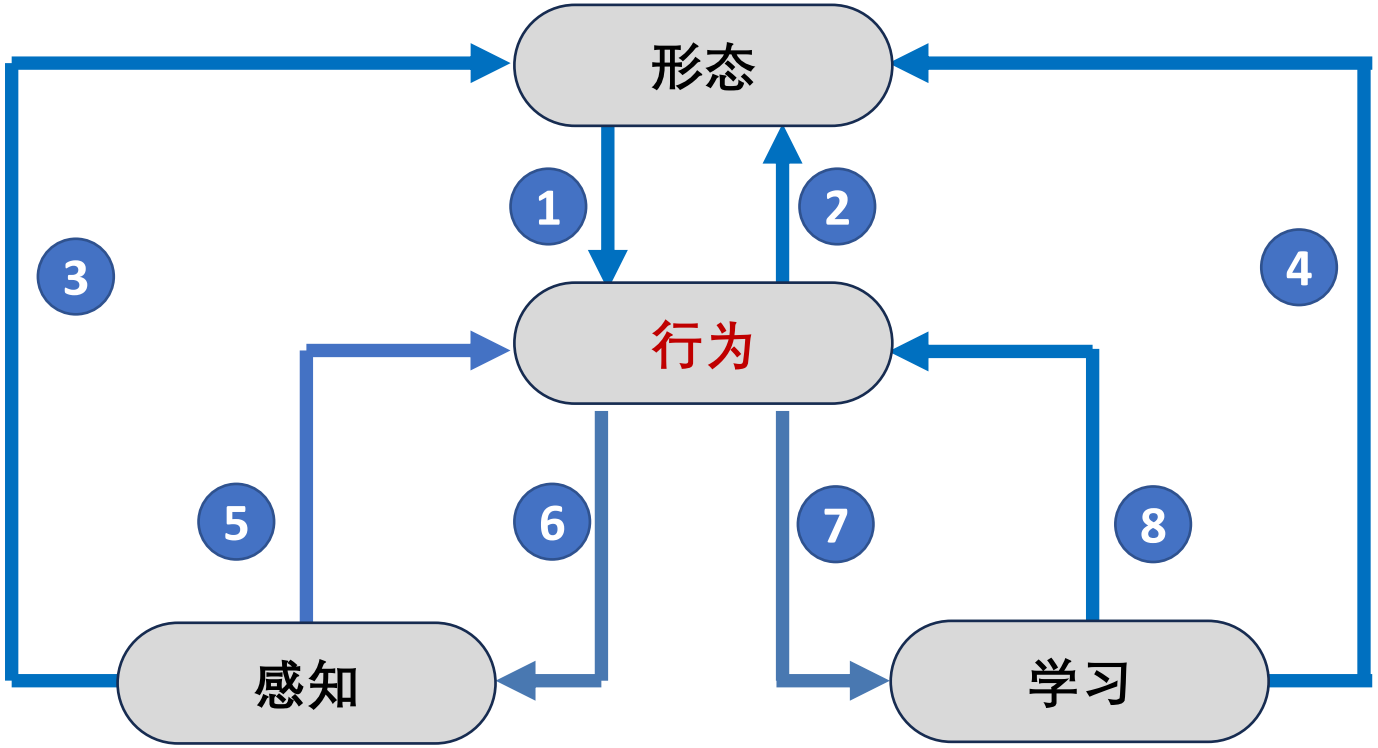
➢ 具身智能的体系结构

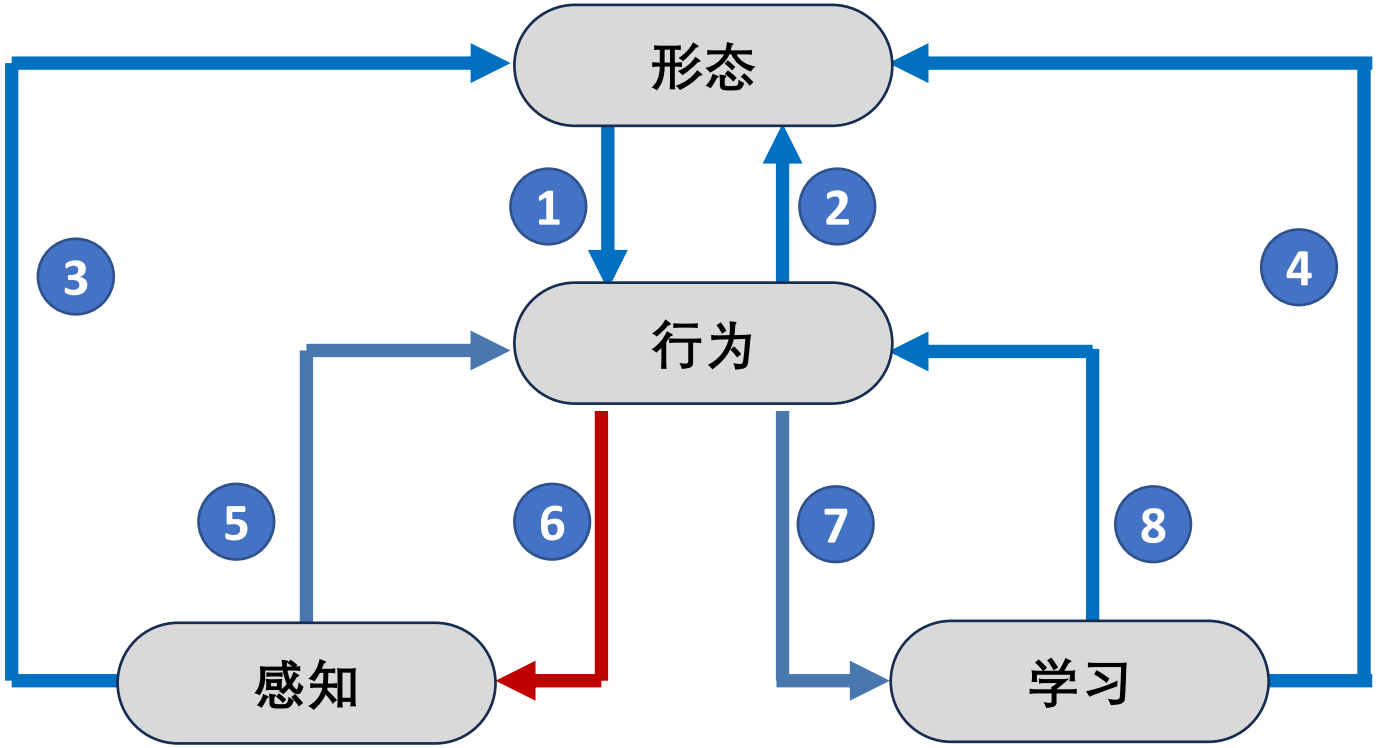




① 基于形态的行为生成　③ 基于感知的形态变换　⑤ 基于感知的行为生成　⑦ 基于行为的自主学习

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


形态
行为
感知
学习
1
2
3
4
5
6
7
8

- **从视觉导航到主动感知**
- **视觉语义导航**
- **具身场景描述**
- **具身语言问答**

1 主动感知

➢ 为什么需要主动感知

眼睛看到的事实其实是大脑想让我们看到的东西

心不在焉

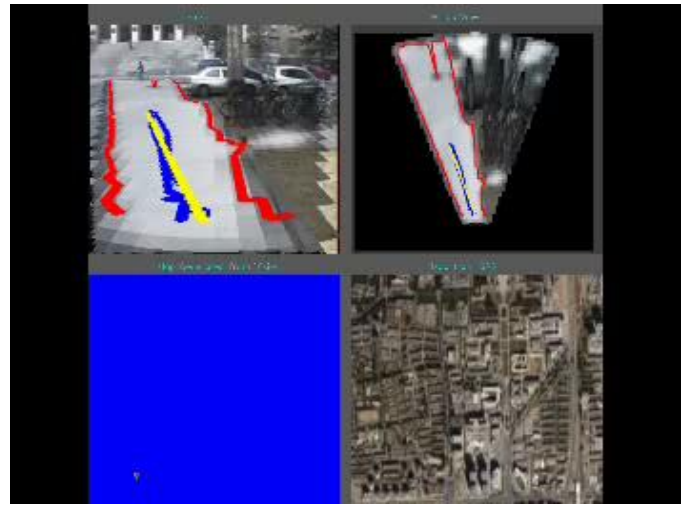
视而不见

听而不闻

食而不知其味

1 主动感知

➢ 为什么需要主动感知



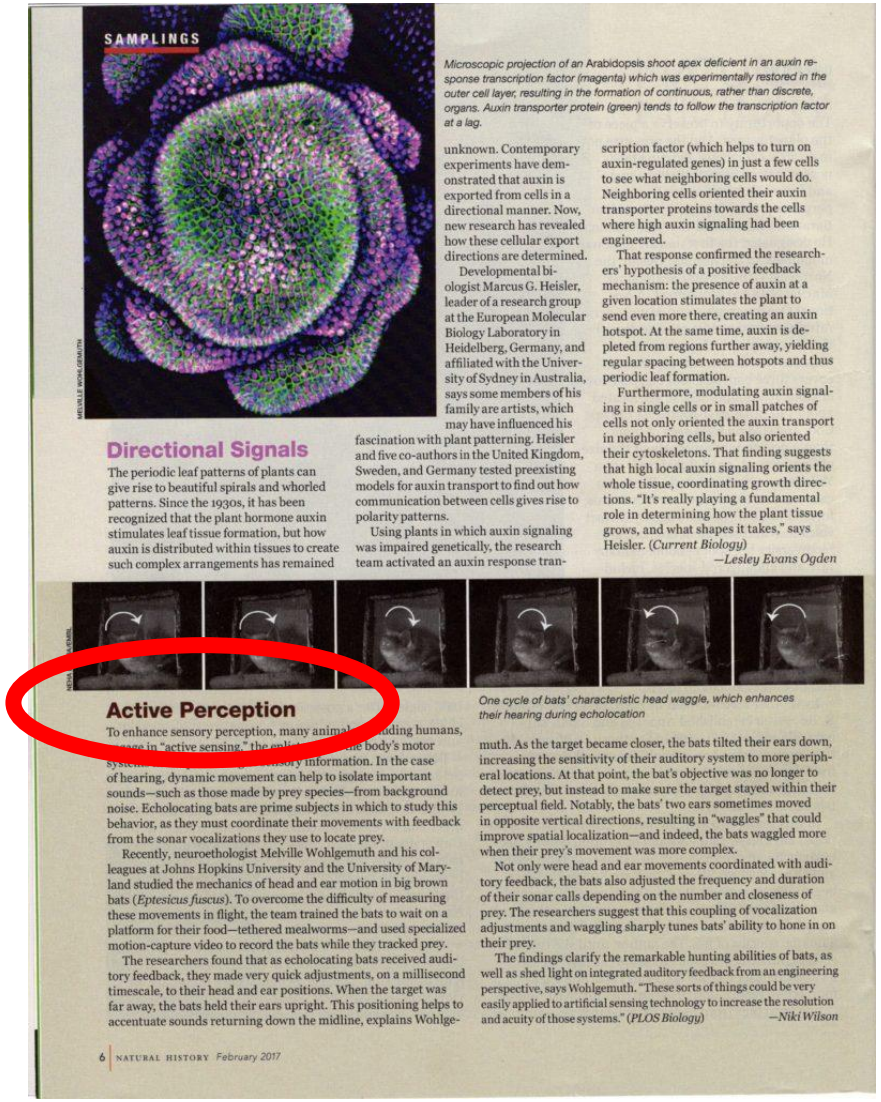
反射式
Reactive

1 主动感知

➢ 生活中的主动感知

1 主动感知

➢ 动物中主动感知



SAMPLINGS

Microscopic projection of an Arabidopsis shoot apex deficient in an auxin response transcription factor (magenta) which was experimentally restored in the outer cell layer, resulting in the formation of continuous, rather than discrete, organs. Auxin transporter protein (green) tends to follow the transcription factor at a lag.

Directional Signals

The periodic leaf patterns of plants can give rise to beautiful spirals and whorled patterns. Since the 1930s, it has been recognized that the plant hormone auxin stimulates leaf tissue formation, but how auxin is distributed within tissues to create such complex arrangements has remained unknown. Contemporary experiments have demonstrated that auxin is exported from cells in a directional manner. Now, new research has revealed how these cellular export directions are determined.

Developmental biologist Marcus G. Heisler, leader of a research group at the European Molecular Biology Laboratory in Heidelberg, Germany, and affiliated with the University of Sydney in Australia, says some members of his family are artists, which may have influenced his fascination with plant patterning. Heisler and five co-authors in the United Kingdom, Sweden, and Germany tested preexisting models for auxin transport to find out how communication between cells gives rise to polarity patterns.

Using plants in which auxin signaling was impaired genetically, the research team activated an auxin response transcription factor (which helps to turn on auxin-regulated genes) in just a few cells to see what neighboring cells would do. Neighboring cells oriented their auxin transporter proteins towards the cells where high auxin signaling had been engineered.

That response confirmed the researchers' hypothesis of a positive feedback mechanism: the presence of auxin at a given location stimulates the plant to send even more there, creating an auxin hotspot. At the same time, auxin is depleted from regions further away, yielding regular spacing between hotspots and thus periodic leaf formation.

Furthermore, modulating auxin signaling in single cells or in small patches of cells not only oriented the auxin transport in neighboring cells, but also oriented their cytoskeletons. That finding suggests that high local auxin signaling orients the whole tissue, coordinating growth directions. "It's really playing a fundamental role in determining how the plant tissue grows, and what shapes it takes," says Heisler. (*Current Biology*)

—Lesley Evans Ogden

One cycle of bats' characteristic head waggle, which enhances their hearing during echolocation

Active Perception

To enhance sensory perception, many animals, including humans, engage in "active sensing," the coordination of the body's motor system with sensory information. In the case of hearing, dynamic movement can help to isolate important sounds—such as those made by prey species—from background noise. Echolocating bats are prime subjects in which to study this behavior, as they must coordinate their movements with feedback from the sonar vocalizations they use to locate prey.

Recently, neuroethologist Melville Wohlgemuth and his colleagues at Johns Hopkins University and the University of Maryland studied the mechanics of head and ear motion in big brown bats (*Eptesicus fuscus*). To overcome the difficulty of measuring these movements in flight, the team trained the bats to wait on a platform for their food—tethered mealworms—and used specialized motion-capture video to record the bats while they tracked prey.

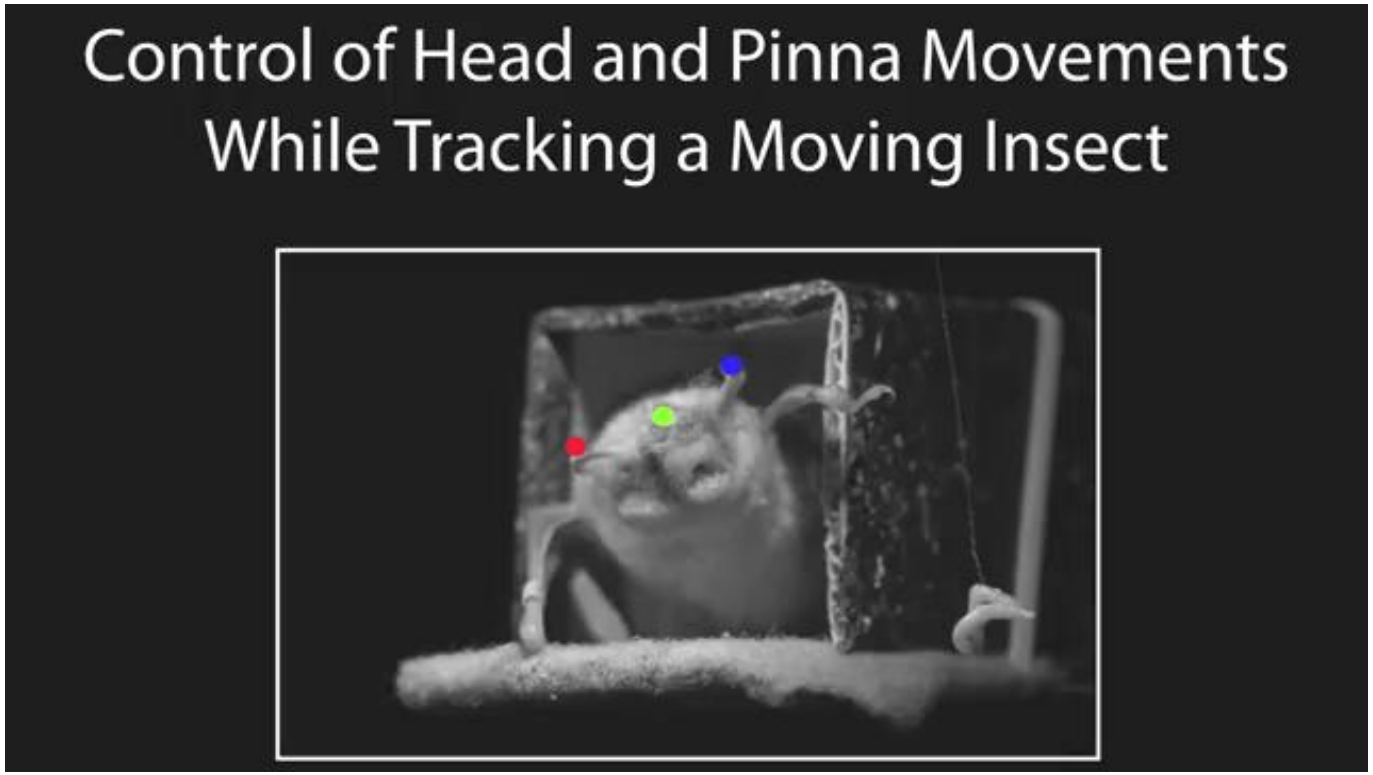
The researchers found that as echolocating bats received auditory feedback, they made very quick adjustments, on a millisecond timescale, to their head and ear positions. When the target was far away, the bats held their ears upright. This positioning helps to accentuate sounds returning down the midline, explains Wohlgemuth. As the target became closer, the bats tilted their ears down, increasing the sensitivity of their auditory system to more peripheral locations. At that point, the bat's objective was no longer to detect prey, but instead to make sure the target stayed within their perceptual field. Notably, the bats' two ears sometimes moved in opposite vertical directions, resulting in "waggles" that could improve spatial localization—and indeed, the bats waggled more when their prey's movement was more complex.

Not only were head and ear movements coordinated with auditory feedback, the bats also adjusted the frequency and duration of their sonar calls depending on the number and closeness of prey. The researchers suggest that this coupling of vocalization adjustments and waggling sharply tunes bats' ability to hone in on their prey.

The findings clarify the remarkable hunting abilities of bats, as well as shed light on integrated auditory feedback from an engineering perspective, says Wohlgemuth. "These sorts of things could be very easily applied to artificial sensing technology to increase the resolution and acuity of those systems." (*PLOS Biology*)

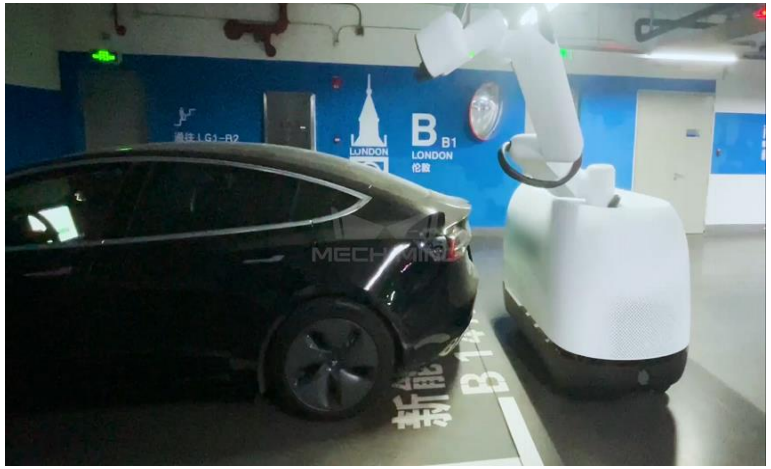
—Niki Wilson

Not only were head and ear movements coordinated with auditory feedback, the bats also adjusted the frequency and duration of their sonar calls depending on the number and closeness of prey.



1 主动感知

➢ 主动感知的应用



1 主动感知

➢ 主动感知的优点

- A motivation for examining active vision is the fact that passive vision has been shown to be very problematic.
- Almost every basic problem in passive machine perception is very difficult, it is ill-posed in the sense of Hadamard.
- Problems that are ill-posed, nonlinear or unstable for a passive observer become well-posed, linear or stable for an active observer.

Active Perception

RUZENA BAJCSY, MEMBER, IEEE

Invited Paper

Active Perception (Active Vision specifically) is defined as a study of Modeling and Control strategies for perception. By modeling we mean models of sensors, processing modules and their interaction. We distinguish local models from global models by their extent of application in space and time. The local models represent procedures and parameters such as optical distortions of the lens, focal lens, spatial resolution, band-pass filter, etc. The global models on the other hand characterize the overall performance and make predictions on how the individual modules interact. The control strategies are formulated as a search of such sequence of steps that would minimize a loss function while one is seeking the most information. Examples are shown as the existence proof of the proposed theory on obtaining range from focus and stereovergence on 2-D segmentation of an image and 3-D shape parametrization.

I. Introduction

Most past and present work in machine perception has involved extensive static analysis of passively sampled data. However, it should be axiomatic that perception is not passive, but active. Perceptual activity is exploratory, probing, searching; percepts do not simply fall onto sensors as rain falls onto ground. We do not just see, we look. And in the course, our pupils adjust to the level of illumination, our eyes bring the world into sharp focus, our eyes converge or diverge, we move our heads or change our position to get a better view of something, and sometimes we even put on spectacles. This adaptiveness is crucial for survival in an uncertain and generally unfriendly world, as millenia of experiments with different perceptual organizations have clearly demonstrated. Yet no adequate account or theory or example of active perception has been presented by machine perception research. This lack is the motivation for this paper.

Manuscript received November 23, 1987; revised March 21, 1988. This work was supported in part by NSF Grant DCR-8410771, Air Force Grant AFOSR F49620-85-K-0018, Army/DAAC-29-84-K-0061, NSF-CER/DCR82-19196 AO2, DARPA/ONR NIH Grant N5-10939-11 as part of Cerebo Vascular Research Center, NIH 1-RO1-NS-23636-01, NSF INT85-14199, NSF DMC85-17315, ARPA N0014-85-K-0807, NATO Grant 0224/85, and by DEC Corporation, IBM Corporation, and LORD Corporation.

The author is with the Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA 19104, USA.

IEEE Log Number 8822793.

II. What is Active Sensing?

In the robotics and computer vision literature, the term "active sensor" generally refers to a sensor that transmits (generally electromagnetic radiation, e.g., radar, sonar, ultrasound, microwaves and collimated light) into the environment and receives and measures the reflected signals. We believe that the use of active sensors is not a necessary condition on active sensing, and that sensing can be performed with passive sensors (that only receive, and do not emit, information), employed actively. Here we use the term active not to denote a time-of-flight sensor, but to denote a passive sensor employed in an active fashion, purposefully changing the sensor's state parameters according to sensing strategies.

Hence the problem of Active Sensing can be stated as a problem of controlling strategies applied to the data acquisition process which will depend on the current state of the data interpretation and the goal or the task of the process. The question may be asked, "Is Active Sensing only an application of Control Theory?" Our answer is: "No, at least not in its simple version." Here is why:

1) The feedback is performed not only on sensory data but on complex processed sensory data, i.e., various extracted features, including relational features.
2) The feedback is dependent on *a priori* knowledge—models that are a mixture of numeric/parametric and symbolic information.

But one can say that Active Sensing is an application of intelligent control theory which includes reasoning, decision making, and control. This approach has been eloquently stated by Tenenbaum [1]: "Because of the inherent limitation of a single image, the acquisition of information should be treated as an integral part of the perceptual process . . . Accommodation attacks the fundamental limitation of image inadequacy rather than the secondary problems caused by it." Although he uses the term *accommodation* rather than *active sensing* the message is the same.

The implications of the active sensing approach are the following:

1) The necessity of models of sensors. This is to say, first, the model of the physics of sensors as well as the noise of the sensors. Second, the model of the signal processing and

0018-9219/88/0800-0996$01.00 © 1988 IEEE

Ruzena Bajcsy

UC Berkeley

John Aloimonos

U. Maryland

Problem	Passive Observer	Active Observer
Shape for shading	Ill-posed problem. Needs to be regularized. Even then, unique solution is not guaranteed because of nonlinearity.	Well-posed problem. Unique solution. Linear equation used. Stability.
Shape from contour	Ill-posed problem. Has not been regularized up to now in the Tichonov sense. Solvable under restrictive assumptions.	Well-posed problem. Unique solution for both monocular or binocular observer.
Shape from texture	Ill-posed problem. Needs some assumption about the texture.	Well-posed problem. No assumption required.
Structure from motion	Well posed but unstable. Nonlinear constraints.	Well posed and stable. Quadratic constraints, simple solution methods, stability.
Optic flow (area based)	Ill posed. Needs to be regularized. The introduced smoothness might produce erroneous results.	Well-posed problem. Unique solution. Might be unstable.

很多困难的视觉感知问题，如果允许“动一动”，其难度会显著降低。

Revisiting active perception

Ruzena Bajcsy[1] · Yiannis Aloimonos[2] · John K. Tsotsos[3]

Auton Robot (2018) 42:177–196
https://doi.org/10.1007/s10514-017-9615-3

Received: 9 September 2016 / Accepted: 11 January 2017 / Published online: 15 February 2017

Abstract Despite the recent successes in robotics, artificial intelligence and computer vision, a complete artificial agent necessarily must include active perception. A multitude of ideas and methods for how to accomplish this have already appeared in the past, their broader utility perhaps impeded by insufficient computational power or costly hardware. The history of these ideas, perhaps selective due to our perspectives, is presented with the goal of organizing the past literature and highlighting the seminal contributions. We argue that those contributions are as relevant today as they were decades ago and, with the state of modern computational tools, are poised to find new life in the robotic perception systems of the next decade.

Keywords Sensing · Perception · Attention · Control

1 Introduction

Some evening, long ago, our ancestors looked up at the night sky, just as they had done for thousands of years. But this time, it was different. For the first time, human eyes noticed patterns in the stars. They saw caricatures of living things and as they scanned the heavens more appeared. These were eventually named, once language permitted, and we now know them as constellations of stars. This perhaps represents the first instance of *connecting the dots*, in a literal as well as figurative manner. How did this come about? What is needed to enable it? That one needs the ability to find a dot, and link one dot to another creating a group, is clear. Underlying this, however, is the critical ability to hypothesize a pattern: that group of stars looks like something I know—let me see if I can find other stars to complete the pattern. This is the essence of active perception—to set up a goal based on some current belief about the world and to put in motion the actions that may achieve it.

Through the years, the topic of perception, and particularly vision, has been a great source of wonder and study by philosophers and scientists alike. This history cannot be reviewed here and the interested reader should see Pastore (1971) and Wade and Wade (2000), among others. This paper presents the history of the computational perspective on the problem of active perception, with an emphasis on visual perception, but broadly applicable to other sensing modalities. Those interested in a biological perspective on active perception should see Findlay and Gilchrist (2003) and Poletti and Rucci (2013). What follows is a brief and selective history of the birth of the computational active perception paradigm. However, there is one source of motivation for the topic that deserves mention. The early work of Gibson (1950) proposed that perception is due to the combination of the environment in which an agent exists and how that agent interacts with the environment. He was primarily interested in optic flow that is generated on the retina when moving through the environment (as when flying) realizing that it was the path of motion itself that enabled the perception of specific elements, while dis-enabling others. That path of motion was under the

This is one of several papers published in *Autonomous Robots* comprising the Special Issue on Active Perception.

All authors contributed equally to all aspects of this manuscript.

✉ John K. Tsotsos
tsotsos@cse.yorku.ca

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[2] Department of Computer Science, University of Maryland, College Park, MD, USA
[3] Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

- Aloimonos, John, Isaac Weiss, and Amit Bandyopadhyay. "Active vision." International journal of computer vision 1, no. 4 (1988): 333-356.

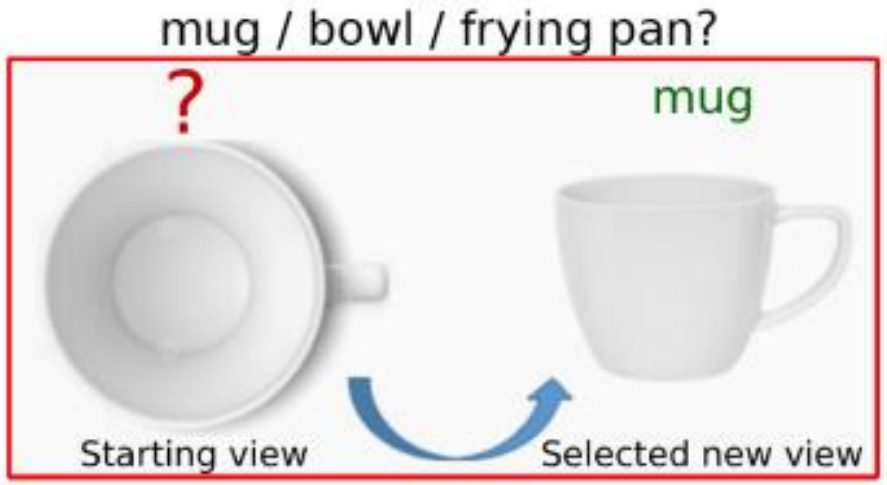
1 主动感知

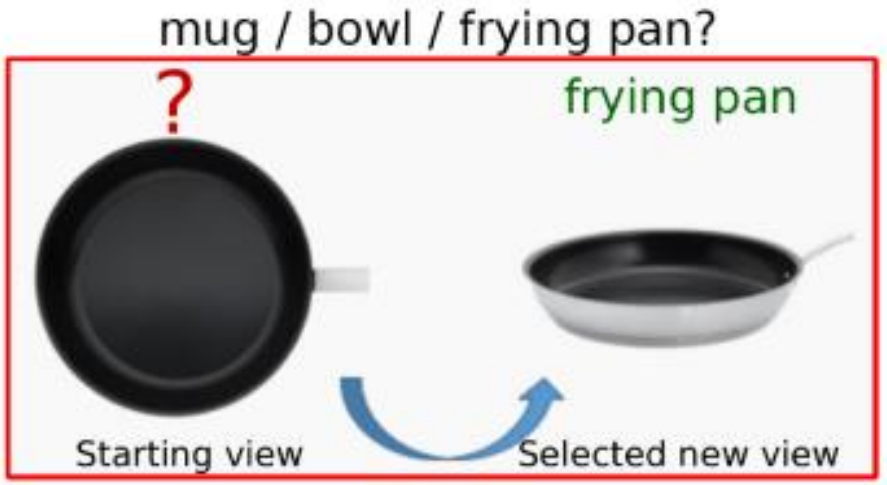
➢ 主动感知的优点

1 主动感知

➢ 主动感知的优点

很多困难的视觉感知问题，如果允许“动一动”，其难度会显著降低。



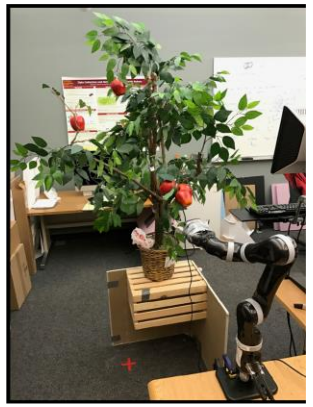


1 主动感知

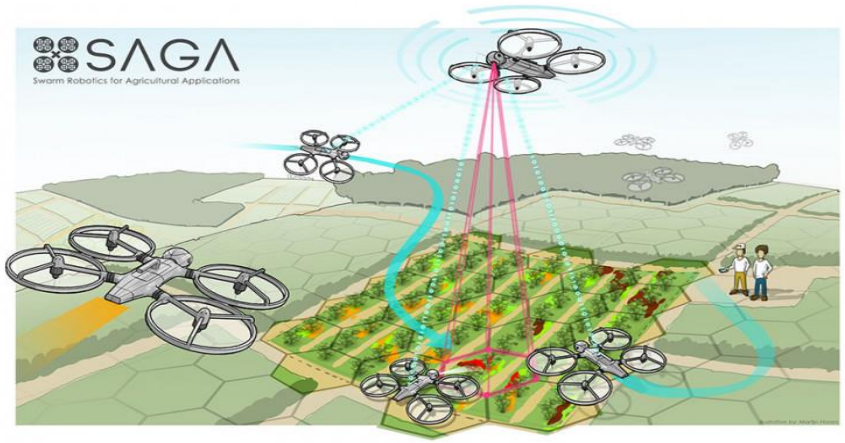
➢ 主动感知的应用




A single view of a cluster can be deceiving (First three images from left). The first view suggests that there are three apples, the second one suggests five and finally from a bottom up view we see that there are actually six apples in the cluster. The rightmost image shows our platform for automating different precision agriculture and phenomics operation

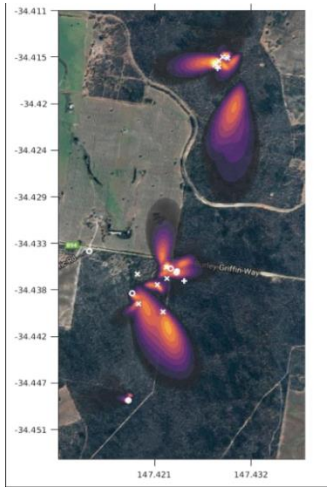



(a) Experimental Setup (b) Planned Views. None of them shows all three apples clearly.



SAGA项目：农业机器人集群
Swarm Robotics for Agricultural Applications


无人机跟踪高速无规律运动且运动范围大的小型动物（雨燕鹦鹉）。



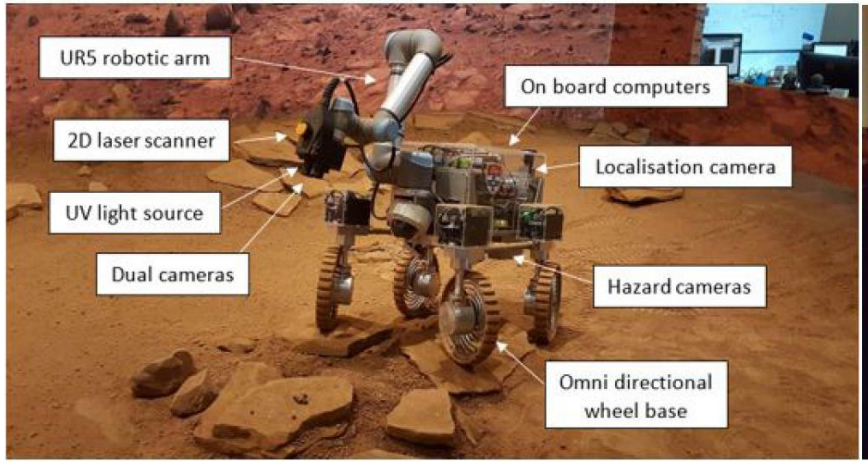
1 主动感知

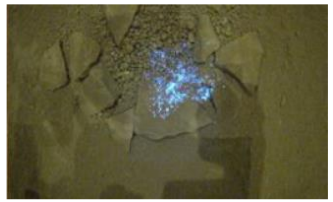
➢ 主动感知的应用

两种模态

Camera

Spectrometer





1 主动感知

➢ 主动感知的关键

- Single view上的感知
  - 常规认为，应该在single view上做到尽可能好。这种想法其实割裂了感知与动作的关系。
  - 其实未必需要这样。有时候在single view上做到最好，需要很大的代价（例如大数据的深度学习训练）。
  - 总结两种模式：“弱感知”+“强控制”； “强感知”+“弱控制”

- 多个view上的融合（证据积累）
  - 常规方法解决的比较多的是同构的，例如，摄像机转来转去，其实获得的都是图像
  - 实际中应该更多地考虑异构的多模态情形
    与常规多模态融合不同的是，主动感知中的融合一般是序贯融合

- 动作生成
  - 基于学习的方法，应该考虑训练时候的动作集与测试时的动作集有差异（比如训练时没有动作约束，但测试时有约束）的问题。这方面的工作在特征表示方面开展的多，但在动作方面很少

- 从视觉导航到主动感知
- 视觉语义导航
- 具身场景描述
- 具身语言问答

2 视觉语义导航

➢ 导航分类

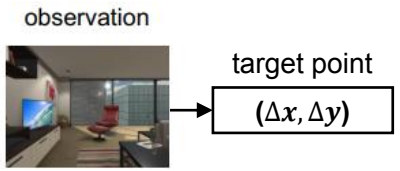
- 视觉环境中常见的三类导航任务：

  1）PointGoal：给定目标坐标点进行导航；

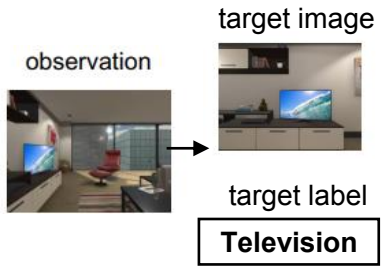
  2）ObjectGoal：给定目标物体的类别，使智能体自主导航至特定物体旁；

  3）AreaGoal：给定目标区域的类别如厨房，使智能体自主导航至特定场景内。

视觉语义导航对应其中的ObjectGoal问题，在视觉语义导航中不考虑PointGoal任务和AreaGoal任务。



PointGoal



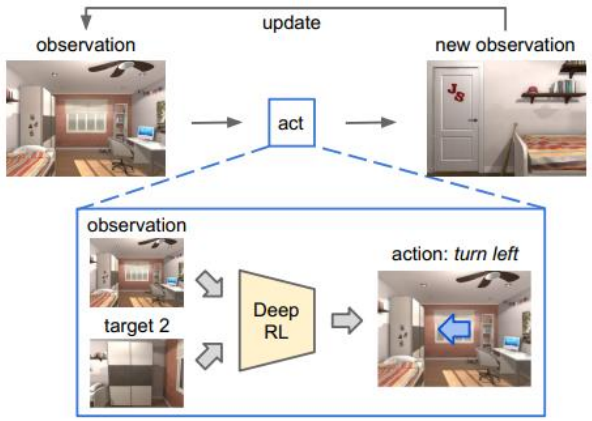
ObjectGoal

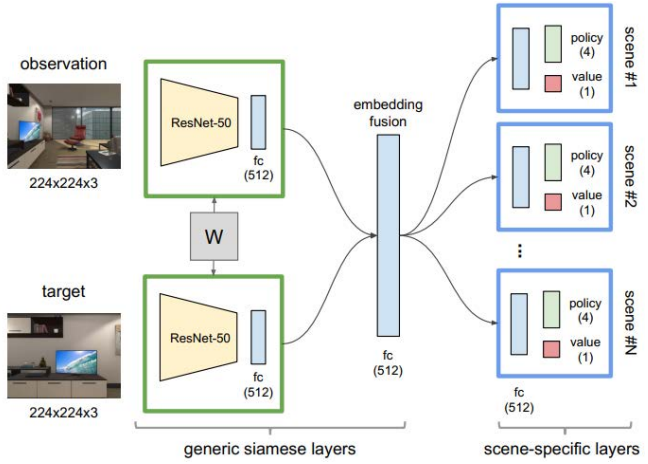
AreaGoal

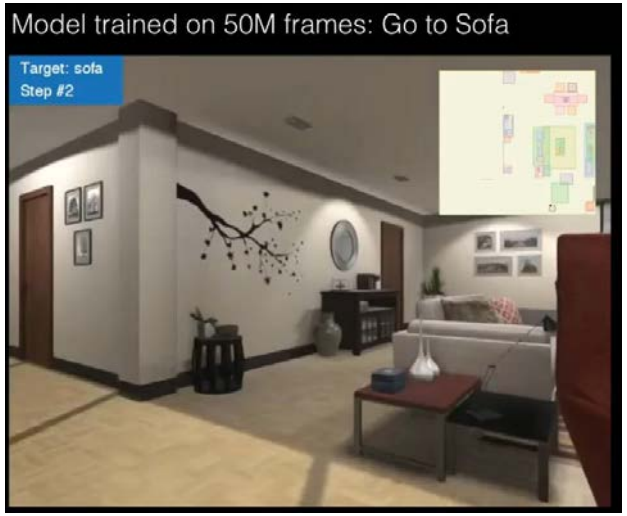
2 视觉语义导航

➢ 目标驱动的导航问题的提出

- 提出室内场景下目标驱动（target-driven）的导航问题，即已知目标的图像信息，在室内场景中智能体仅使用第一视角图像信息通过与环境的交互完成对目标的导航任务。
- 提出一种孪生网络的基本结构，是最早的端到端解决视觉导航问题的模型，成为目标驱动导航问题模型的基本结构。







- Yuke Zhu, et al. Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning. ICRA 2017

2 视觉语义导航

➢ 目标驱动的导航问题的提出

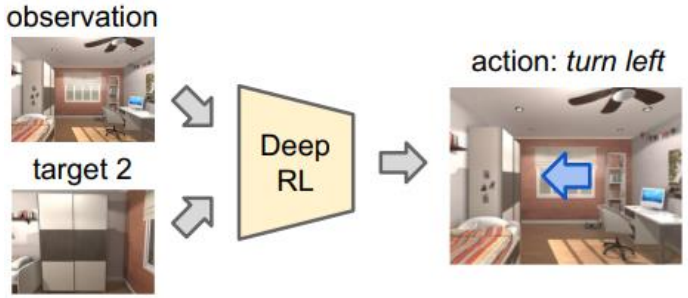
按其已知的输入信息将其分为两类：

1）已知目标物体的**图像**；

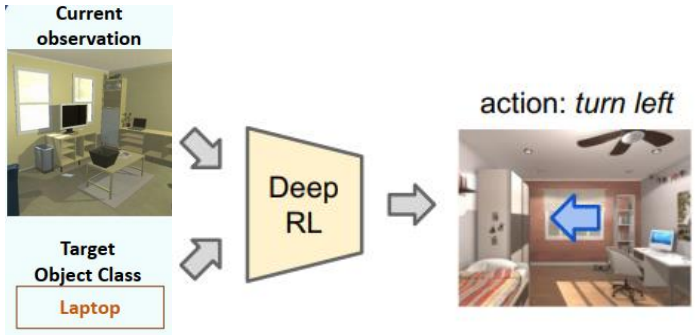
2）已知目标物体的**语义或类别标签**

一般所指的标准的视觉语义导航是第二类问题，即仅已知目标物体的类别，如书、苹果等。

知识图谱、深度图、地图重建等方法被广泛应用于视觉语义导航问题。



目标物体图像



目标物体语义标签

2 视觉语义导航

➢ 分类

- 视觉语义导航：给定目标的语义标签，仅使用视觉输入在环境中导航至特定目标。在视觉语义导航中，不考虑已知场景地图或者已知目标物体位置信息的情况。

- 任务类型：

  已见过场景&已见过目标

  已见过场景&未见过目标

  未见过场景&已见过目标

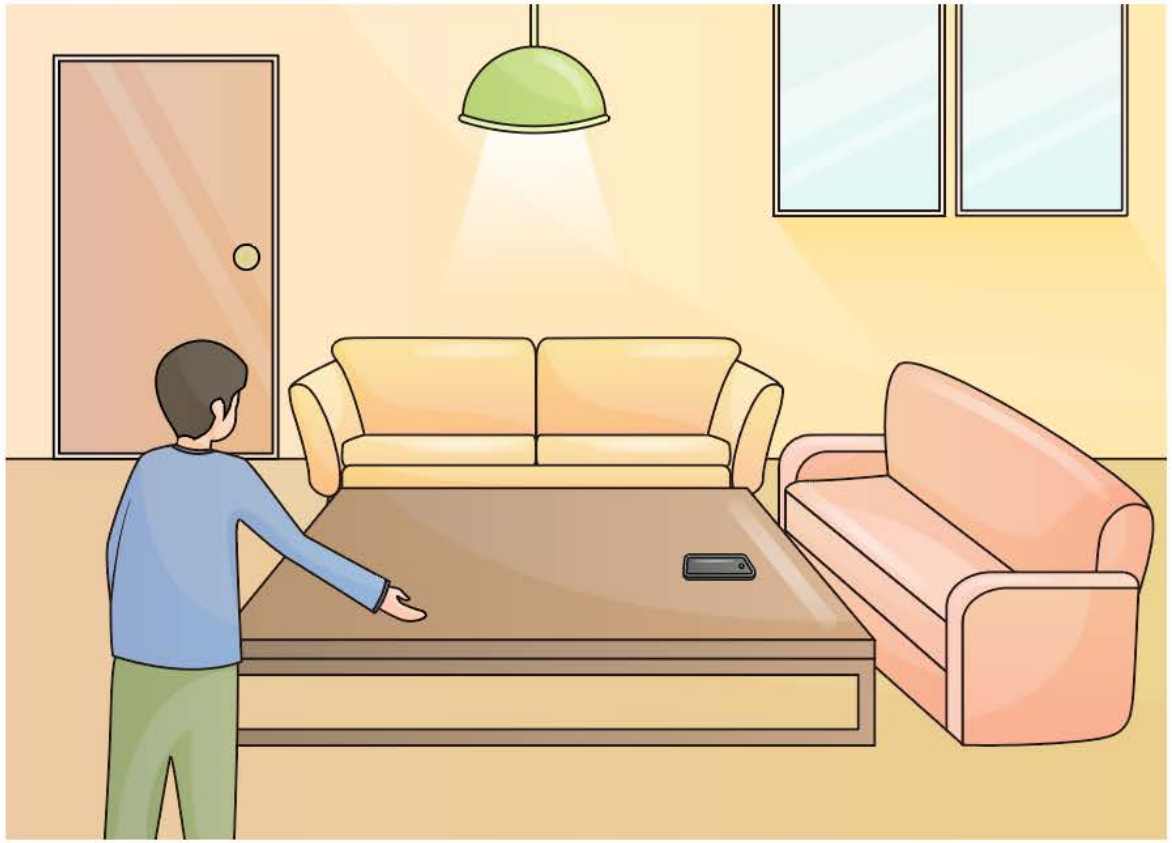
  未见过场景&未见过目标

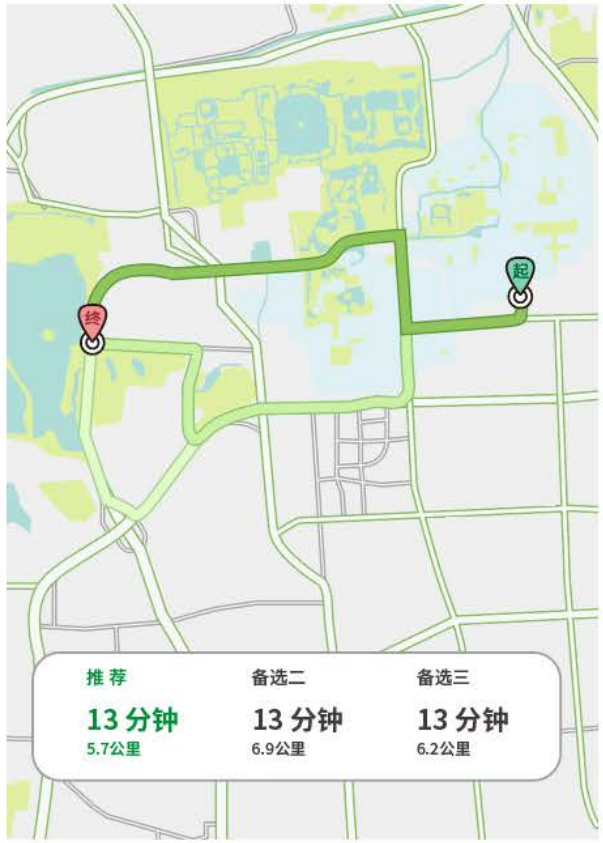
- 常见方法：

  基于学习的导航方法

2 视觉语义导航

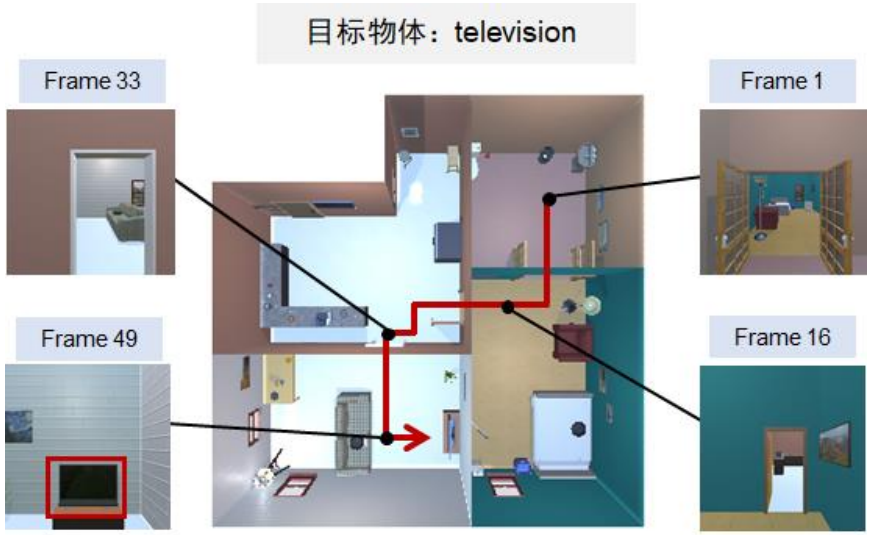
➢ 视觉语义导航（Visual Semantic Navigation, VSN）





2 视觉语义导航

➢ VSN问题描述



$$s_t = \varphi(I_t)$$

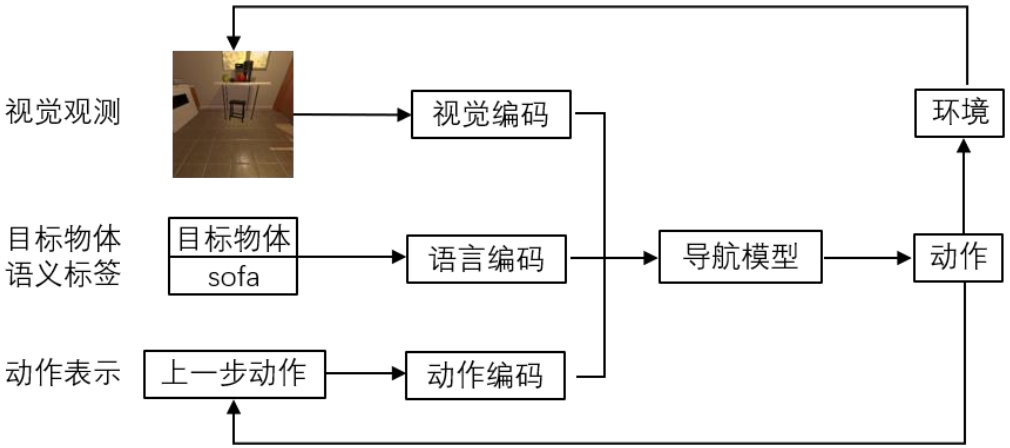
$$o = \phi(O)$$

$$a_t = \mathrm{Nav}(s_t, a_{t-1}, o)$$

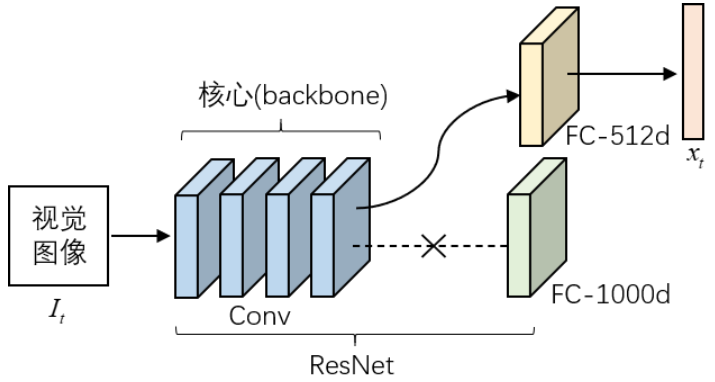
$$s_{t+1} = \mathrm{Env}(s_t, a_t)$$

2 视觉语义导航

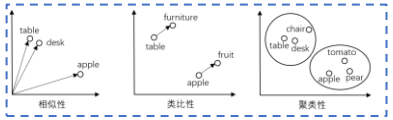
➢ VSN基本方法

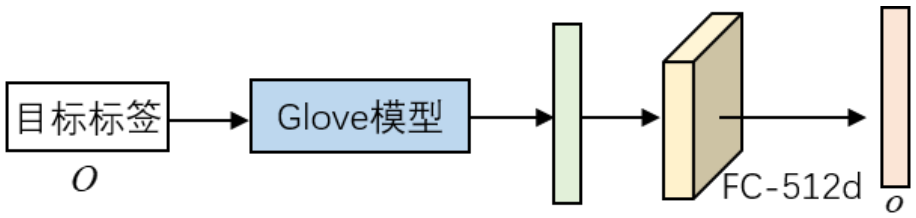


视觉图像特征提取

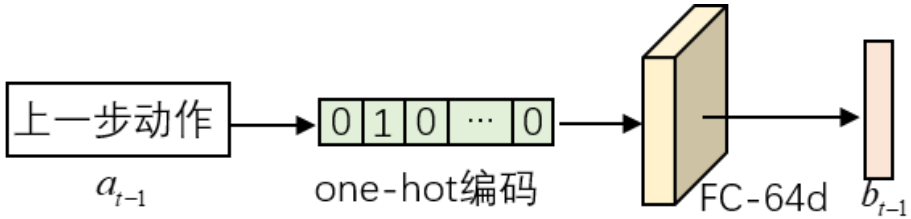


语义特征提取





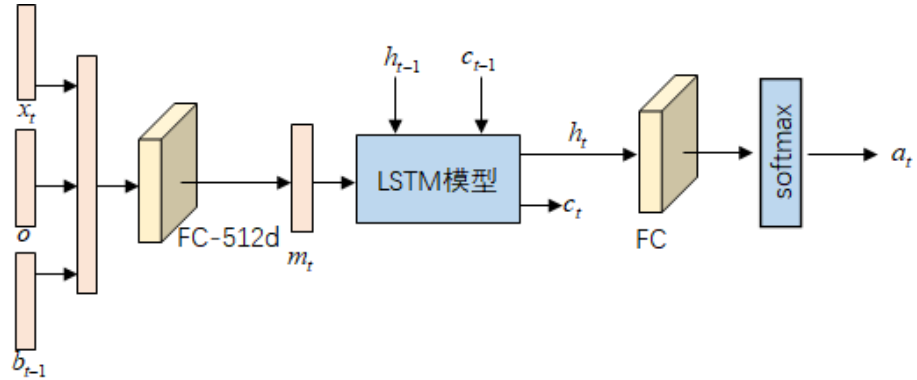
动作特征提取



2 视觉语义导航

➢ VSN基本方法

导航模型

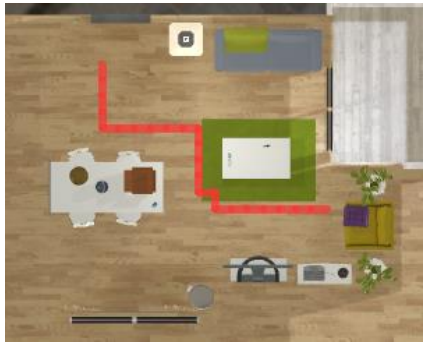


- 强化学习

$$r_t = \begin{cases} 5 & \text{成功找到目标物体} \\ -0.01 & \text{其他} \end{cases}$$

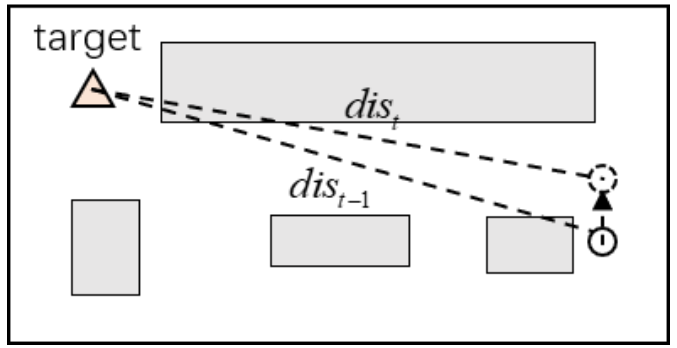
$$r_t = \begin{cases} 5 & \text{成功找到目标物体} \\ dis_t - dis_{t-1} + penalty & \text{其他} \end{cases}$$

- 模仿学习



Target: Pillow

$$L_\theta = \sum_{t=1}^{T} -\log \pi_\theta \left(a_t \mid s_1, a_1, \cdots, s_{t-1}, a_{t-1} \right)$$



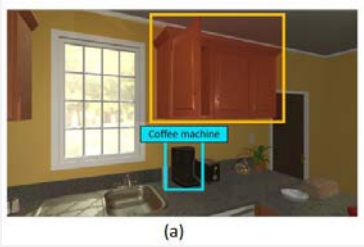
2 视觉语义导航

➢ 融入场景图谱的VSN

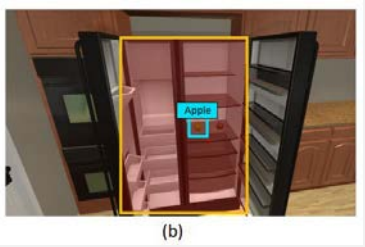
动机

启发

- 寻找马克杯时，当前视角中没有马克杯，但根据以往的经验，马克杯的位置很有可能是在咖啡机旁边的橱柜里。
- 寻找芒果时，智能体之前没有见过芒果，但根据经验，芒果和苹果是类似的物体都是水果，而智能体之前在冰箱中见过苹果，于是推断出芒果有可能的位置是在冰箱里。



(a)

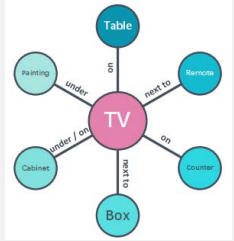


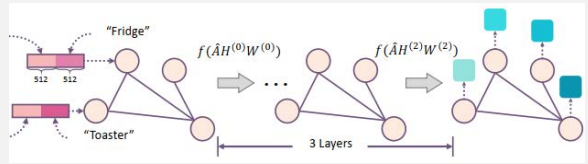
(b)

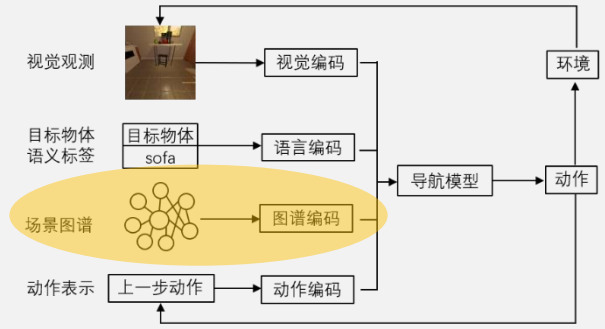
引入场景图谱的优势：

- 有效编码不同类物体之间的空间关系
- 对于智能体未见过的物体类别，场景图谱能够提供其与已见过物体之间的空间和视觉相对关系。

方法







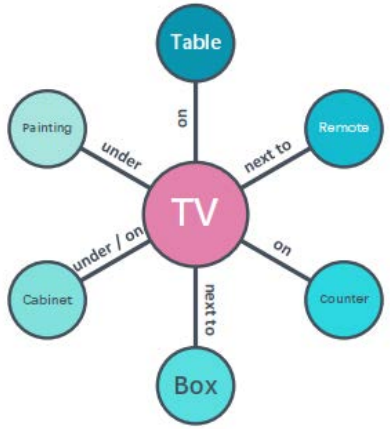
$$a_t = \mathrm{Nav}(s_t, a_{t-1}, o, g)$$

2 视觉语义导航

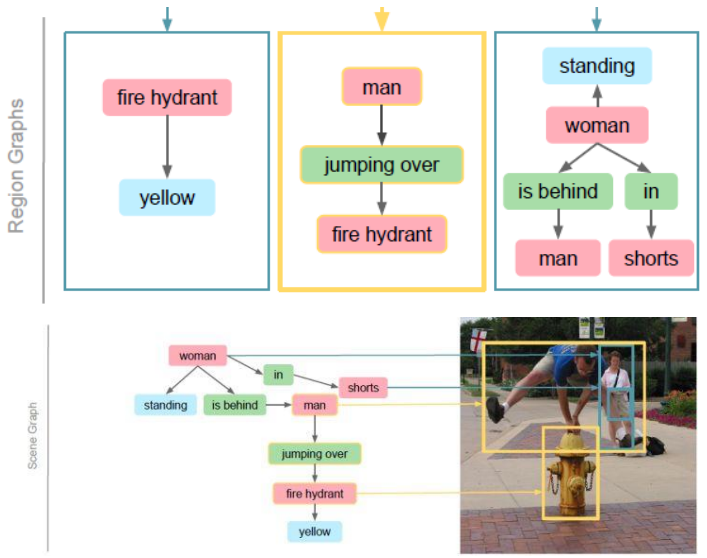
➢ 融入场景图谱的VSN

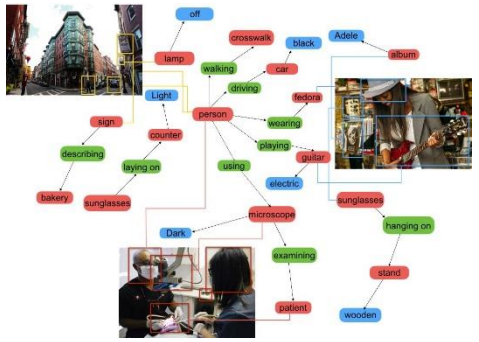
➢ 场景图谱的构建

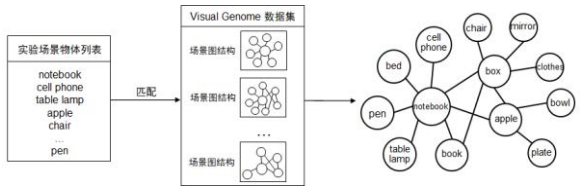
将场景图谱记为$G = (V, E)$，其中V, E分别代表图中的节点和边。每个节点$v \in V$代表一个物体类别，每条边$e \in E$代表一种相对位置关系。使用Visual Genome数据集构建场景图谱，该数据集包含超过十万张图片，每张图片都被标注了对应的物体和物体间的相对关系，其包含相当多的物体种类。



- **关于Visual Genome**
  - Visual Genome（VG）是李飞飞组于2016年发布的大规模标注的图片语义理解数据集。
  - VG数据集中的每张图片，包含Region Description，Region Graph，Scene Graph
    - Region Description：图片被分成若干region，每个region有与其对应的一句自然语言描述
    - Region Graph：提取每个region中的目标、属性和关系构成局部 Graph
    - Scene Graph：把图片中所有Region Graph合并成一个全局 Scene Graph







2 视觉语义导航

➢ 融入场景图谱的VSN

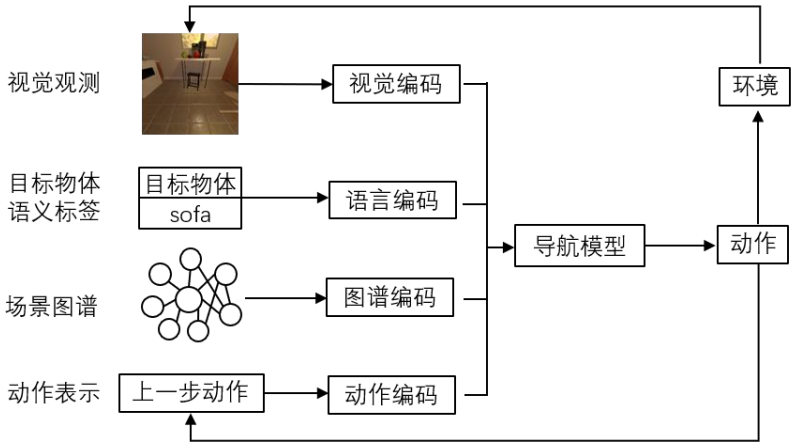
➢ 场景图谱的处理

- **图卷积网络（GCN）**

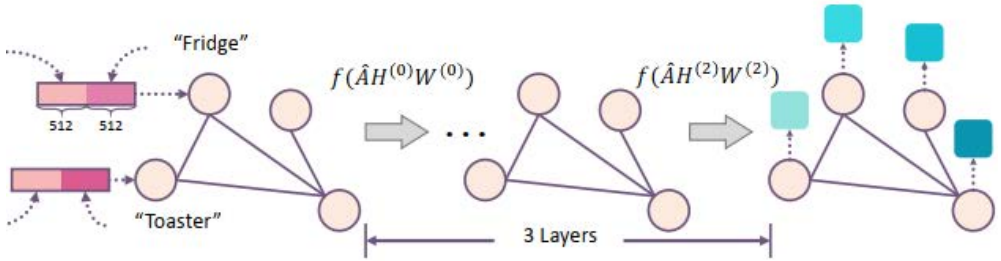
GCN是CNN在图结构下的变种，用于从图结构$G = (V, E)$中学习到一定的函数表达式。每个节点v的输入是一个特征向量x_v。所有节点的输入可表示为一个矩阵$X = [x_1, \ldots, x_{|V|}] \in R^{|V| \times D}$，其中D是输入向量的维度。整个图结构可以表示为一个二进制邻接矩阵A，对其进行正交化得到$\hat{A}$。GCN网络输出一个节点级的向量表示$Z = [z_1, \ldots, z_{|V|}] \in R^{|V| \times F}$。 ReLU激活函数记为$f(\cdot)$，对于GCN网络有

$$H^{(l+1)} = f\left(\hat{A}H^{(l)}W^{(l)}\right)$$

其中$H^{(0)} = X$， $H^{(L)} = Z$， $W^{(l)}$是GCN第l层的权重矩阵，L是GCN网络的层数。



$$a_t = \mathrm{Nav}(s_t, a_{t-1}, o, g)$$



2 视觉语义导航

➢ 融入场景图谱的VSN

➢ 例子

将智能体所处环境进行网格化。动作空间由平移运动和旋转运动构成。

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°。

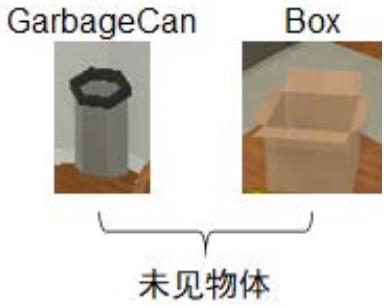
智能体通过执行动作空间中的动作 ，不断对环境进行探索，当智能体执行停止动作 时，结束对环境的探索。

- 随机探索模型：智能体每一步从其动作空间中随机选择探索动作。
- 深度优先探索模型：智能体每一步优先沿着当前可行动作方向向前探索，并优先向右移动。
- IL：利用模仿学习训练动作生成模型。
- IL+A3C：利用模仿学习与强化学习相结合的方式训练动作生成模型。
- IL+A3C+场景图谱：引入场景图谱，采用IL+A3C的方法训练模型。

2 视觉语义导航

➢ 融入场景图谱的VSN

见过目标物体导航定量结果



未见过目标物体导航定量结果

$d \geq 5$ $d \geq 5$

实验设置	模型	准确率(%)	SPL (%)	准确率(%)	SPL(%)
已见过目标物体	随机	8.3	3.66	5.2	2.33
	深度优先探索	15.2	10.12	10.3	8.55
	IL	48.3	18.67	37.1	13.12
	IL+A3C	52.8	20.64	41.3	15.28
	IL+A3C+场景图谱	55.1	21.38	45.7	17.32

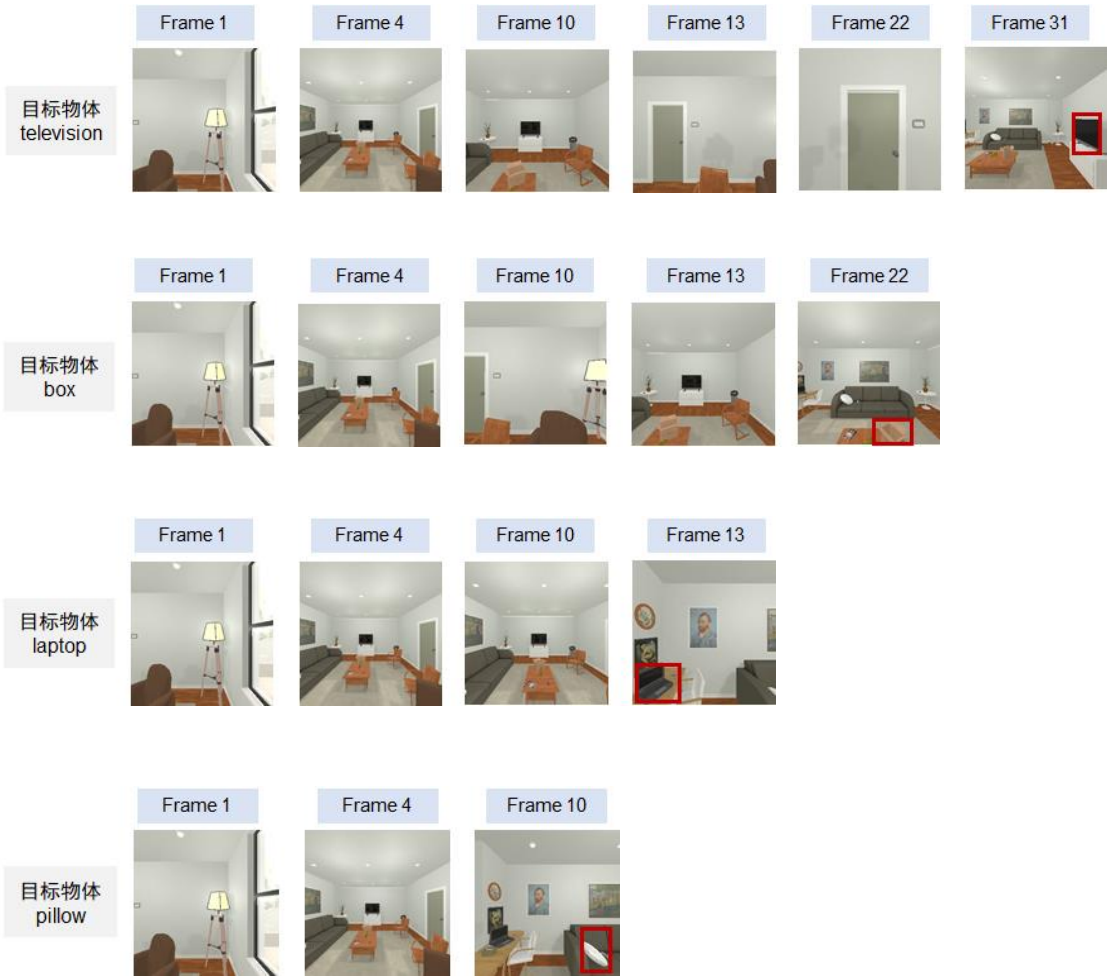
实验设置	模型	准确率(%)	SPL (%)	准确率≥5(%)	SPL≥5(%)
未见过目标物体	随机	8.2	3.58	5.2	2.32
	深度优先探索	15.0	10.11	10.2	8.53
	IL	37.1	13.65	24.3	8.78
	IL+A3C	43.7	15.23	28.1	10.05
	IL+A3C+场景图谱	45.5	17.13	30.3	12.65

2 视觉语义导航

➢ 融入场景图谱的VSN

AI2-THOR

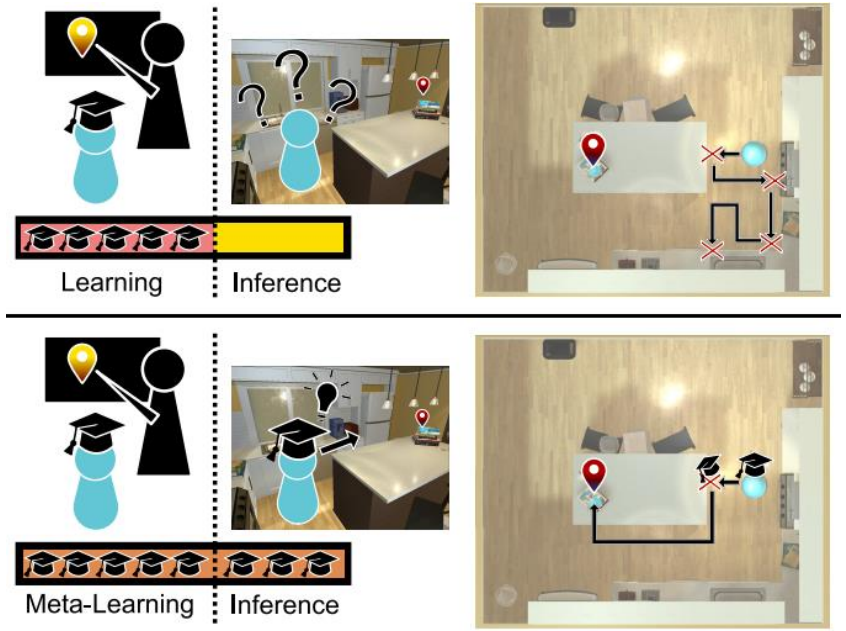
厨房　客厅　卧室　浴室



2 视觉语义导航

➢ 前沿：基于元学习的自适应视觉语义导航

- 传统的强化学习或者深度学习方法，训练结束后执行某项任务时，网络结构及参数固定不变。
- 视觉语义导航任务的难点是智能体对在训练过程中未见过的新场景的泛化能力，因为不同场景中房间结构和物品摆放有较大的区别。为此，提出一种基于元学习的自适应方法，在训练和执行的过程中都在不断调整模型参数，使之能适应新场景。



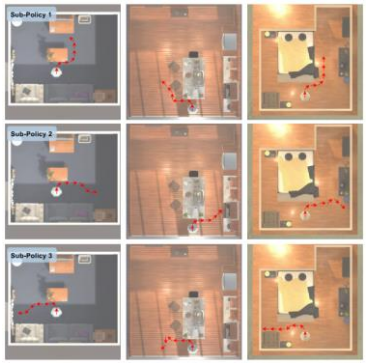
- Learning to Learn How to Learn: Self-Adaptive Visual Navigation using Meta-Learning (CVPR 2019) Mitchell Wortsman1, Kiana Ehsani et al.

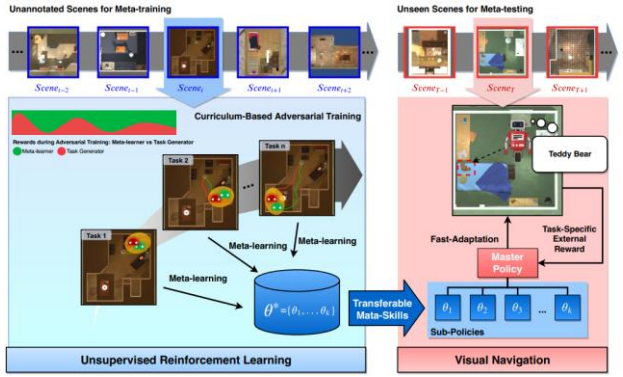
2 视觉语义导航

➢ 前沿：基于无监督学习的视觉语义导航

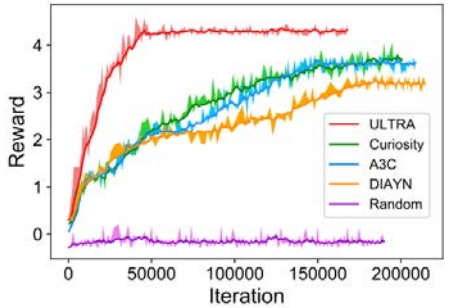
在无监督信号的情况下从环境中学习可迁移的元技能（meta-skills）和子策略（sub-policy）如绕过障碍物等。当新的环境提供相应的外部奖励时，智能体可以通过学习一个高级的master策略来组合之前学习到的元技能，从而快速适应视觉导航任务。

模型结构：蓝色部分是对抗训练过程，任务生成器自动生成越来越难的任务，而元学习模块用于学习如何完成这些任务。 从这些任务中，元学习模块可以学习一组可迁移的子策略。在右侧，给定任务特定的外部奖励函数，元学习模块基于已学习到的可迁移策略通过学习新的master策略可以快速适应视觉导航任务。





	All		$L \geq 5$	
	Success	SPL	Success	SPL
Random	8.21	3.74	0.24	0.09
A3C (learn from scratch)	19.20	7.48	9.43	4.13
DIAYN	17.23	6.30	8.72	3.79
Curiosity	21.07	8.51	10.31	4.37
Ours				
ULTRA	**27.74**	**11.47**	**20.57**	**8.04**
– hierarchical policy	24.27	10.54	14.13	5.61
– adversarial training	20.23	8.35	10.04	4.33

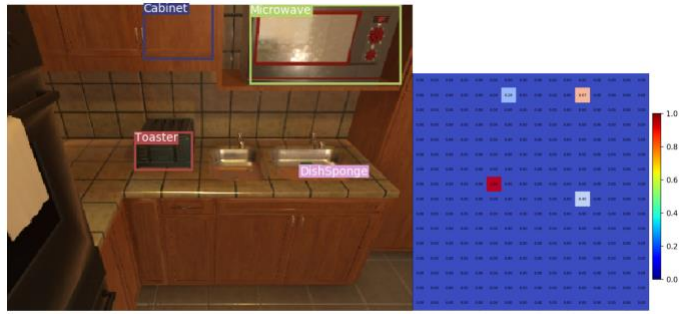


- Unsupervised Reinforcement Learning of Transferable Meta-Skills for Embodied Navigation（CVPR 2020）Juncheng Li, Xin Wang et al.

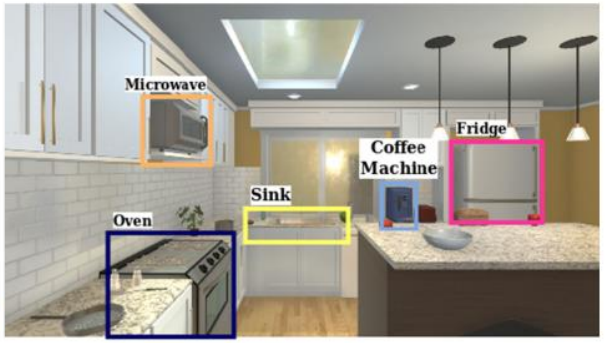
2 视觉语义导航

➢ 前沿：基于上下文信息的视觉语义导航

- 人们寻找目标时不会将所有场景都看一遍，而是寻找最相关的物体。例如，智能体当前在位置(a)寻找海绵(b)，当前位置看不到海绵。利用空间的上下文信息（context）作为线索，如海绵应该在水槽附近，而水槽很容易被检测到，这使得智能体从看不见目标的位置开始也可以导航到目标。
- 提出了一种基于上下文信息的导航方式，使用context grid的表示方式来表示当前场景中的物体与目标物体的相似度，从而帮助机器人找到更小的或隐藏的目标。该方法具有更好的泛化性。



(a) YOLO output (b) Context grid (16 × 16)



(a)

(b)

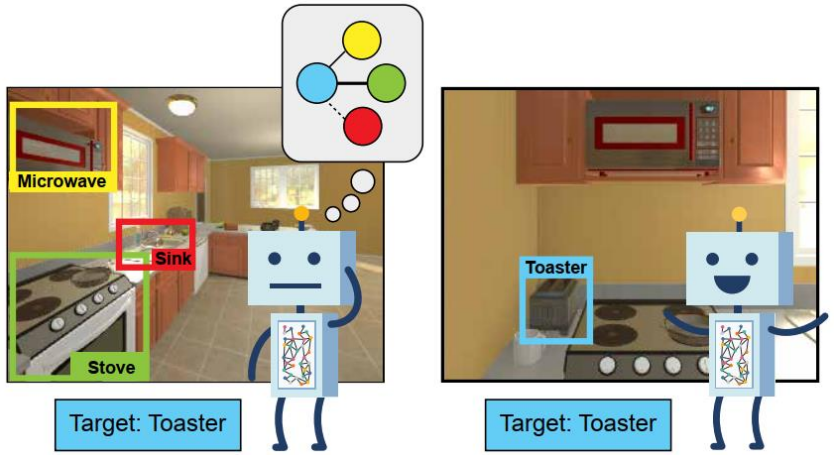
- Visual Object Search by Learning Spatial Context (RA-L 2020). Raphael Druon , Yusuke Yoshiyasu et al.

2 视觉语义导航

➢ 前沿：基于层次关系学习的视觉语义导航

人们在寻找物体时会建立不同物体间的内在层次关系。例如，在厨房中找烤面包机时，人会倾向于先找可能与之相邻的更大的候选物体如微波炉或者炉灶，然后再在这些物体周围寻找烤面包机。

将与目标物体有空间或者语义关系的目标称为parent-object，提出了一种层次物体关系学习方法，在图卷积网络中使用Context Vector作为结点向量表示。通过一种新的reward shaping机制引入层次物体关系。



- Learning hierarchical relationships for object-goal Navigation （CoRL 2020）Yiding Qiu Anwesan Pal & Henrik I. Christensen

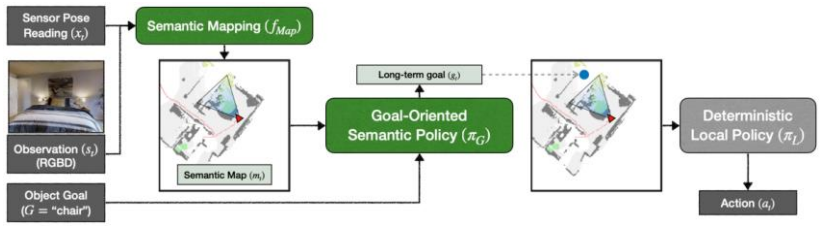
2 视觉语义导航

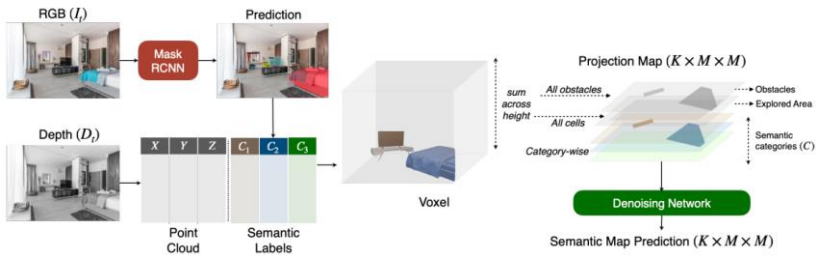
➢ 前沿：基于建图的视觉语义导航

- 在未知场景中进行目标导航时，智能体不仅需要识别目标，还需要知道在哪里更容易找到目标物体。
- 基于建图的导航方法，将第一视角通过Mask-RCNN获取语义信息，映射为Top-down视角的二维地图，利用该地图训练一种导航策略。

模型结构

模型包含两个模块，语义地图构建模块，和基于语义地图的语义决策模块。使用预训练的Mask-RCNN检测，映射到点云中生成带有语义的像素映射，将不同类别的物体映射到不同的地图层中生成语义地图。若当前状态发现目标，语义决策模块将目标作为导航点；若当前没有发现目标，则根据目标类别和当前语义地图，推测最有可能找到的位置作为导航点。





- Object Goal Navigation using Goal-Oriented Semantic Exploration（CVPR 2020）Devendra Singh Chaplot, Dhiraj Gandhi et al.

2 视觉语义导航

➢ 前沿：基于建图的视觉语义导航

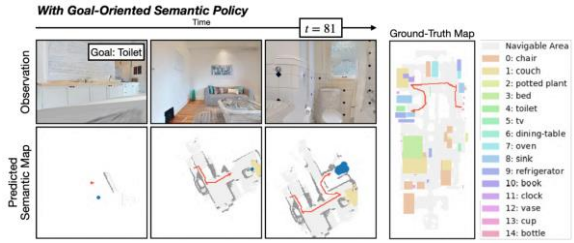
- **实验结果**

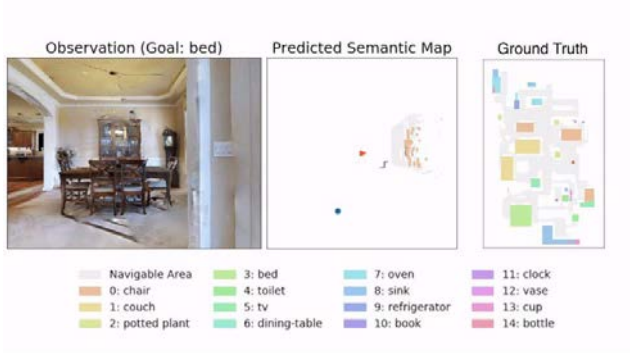
实验平台：Habitat

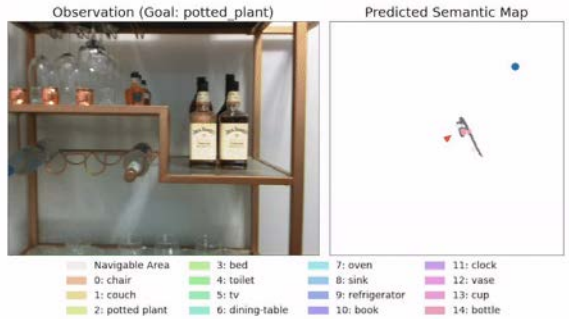
结果：采用语义地图的方法比单纯使用强化学习进行端到端训练的效果在成功率，SPL和DTS（智能体停止时与成功位置的距离）上都好很多。

Table 1: **Results.** Performance of SemExp as compared to the baselines on the Gibson and MP3D datasets.

Method	Gibson SPL	Gibson Success	Gibson DTS (m)	MP3D SPL	MP3D Success	MP3D DTS (m)
Random	0.004	0.004	3.893	0.005	0.005	8.048
RGBD + RL [38]	0.027	0.082	3.310	0.017	0.037	7.654
RGBD + Semantics + RL [31]	0.049	0.159	3.203	0.015	0.031	7.612
Classical Map + FBE [46]	0.124	0.403	2.432	0.117	0.311	7.102
Active Neural SLAM [9]	0.145	0.446	2.275	0.119	0.321	7.056
SemExp	**0.199**	**0.544**	**1.723**	**0.144**	**0.360**	**6.733**



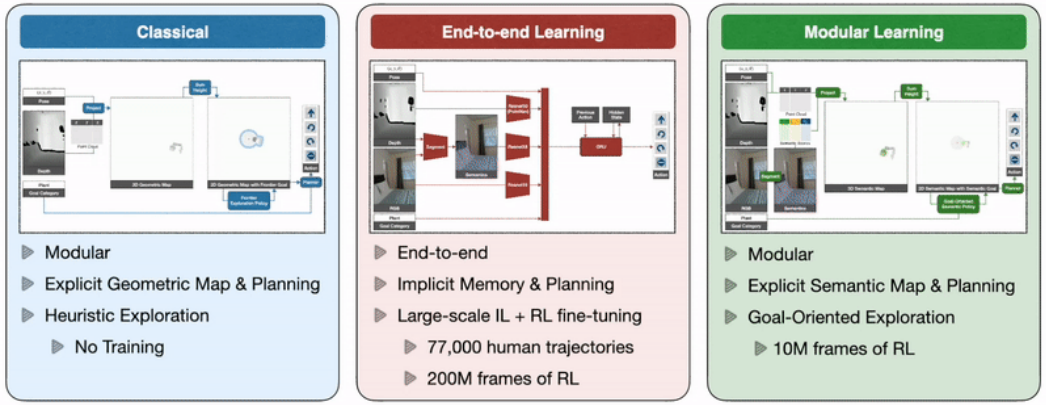


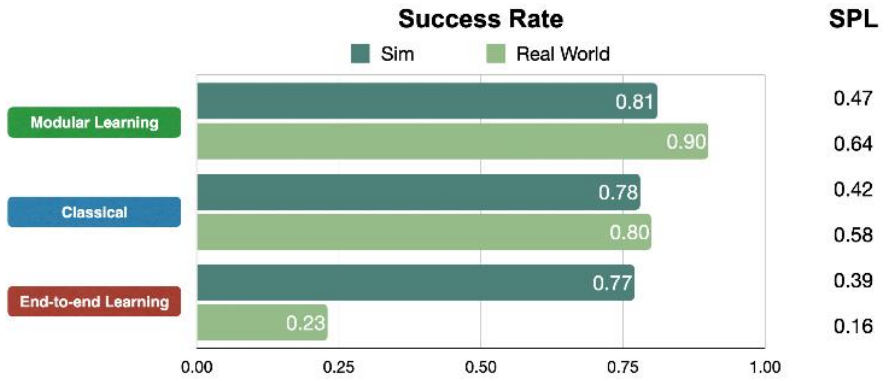


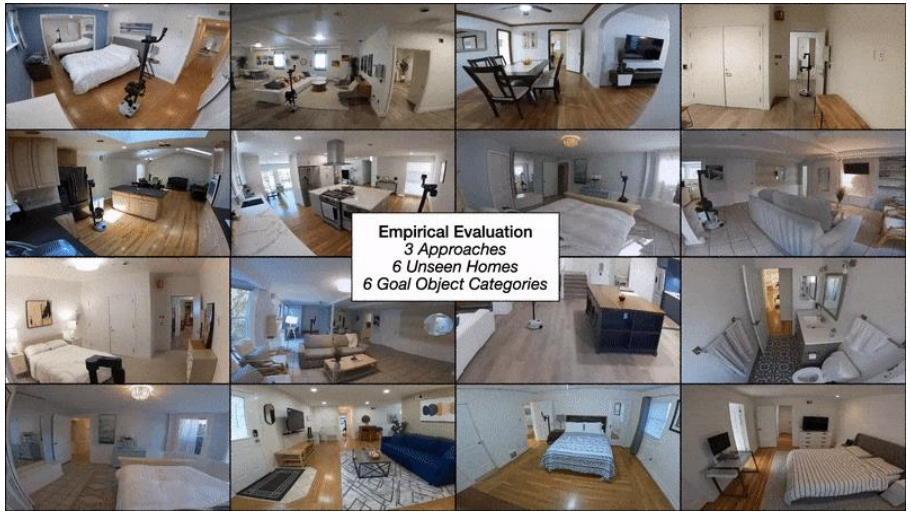
- Object Goal Navigation using Goal-Oriented Semantic Exploration（CVPR 2020）Devendra Singh Chaplot, Dhiraj Gandhi et al.

2 视觉语义导航

➢ 前沿：基于建图的视觉语义导航





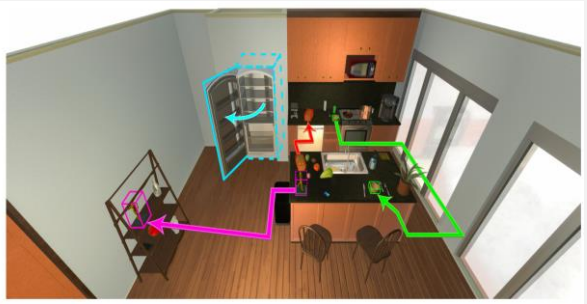


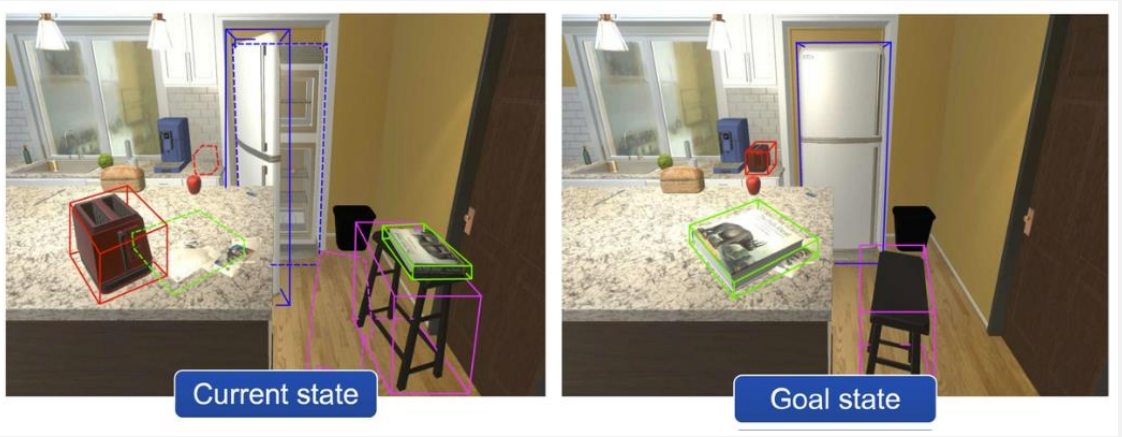
- Theophile Gervet, Soumith Chintal, Dhruv Batra, Jitendra Malik, Devendra Singh Chaplot, Navigating to Objects in the Real World

2 视觉语义导航

➢ 前沿

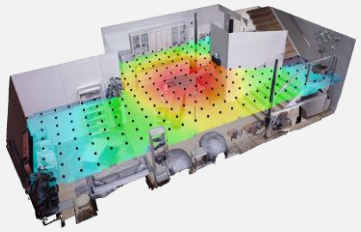
Rearrangement

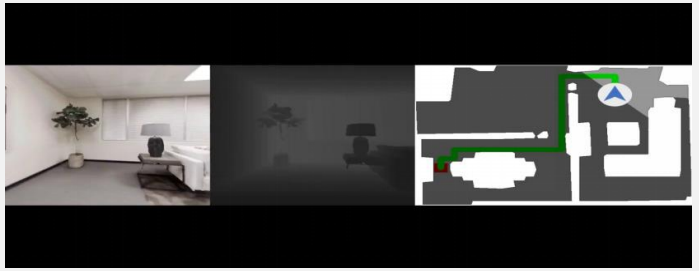




- Weihs, Luca, et al. , Visual room rearrangement. (CVPR 2021).

Audio-Visual Navigation





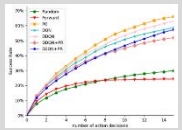
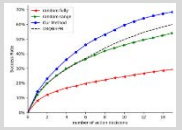
SoundSpaces：一个声音模拟平台，用于实现两个视觉逼真的3D环境（Replica和Matterport3D）的视听导航

- Chen Changan, et al. , Soundspaces: Audio-visual navigation in 3d environments. ECCV 2020

2 视觉语义导航

➢ 前沿

主动目标发现

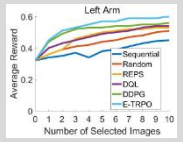




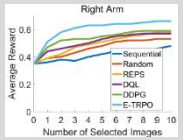
- Active object detection with multi-step action prediction using deep Q-network, IEEE Transactions on Industrial Informatics, 2019

主动目标识别



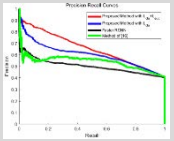
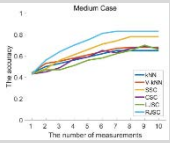






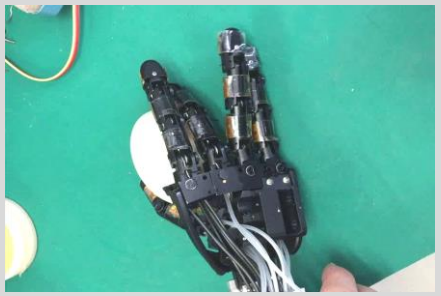
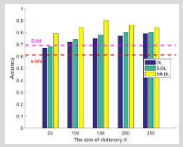
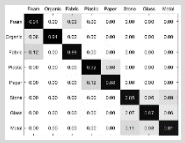
- Extreme trust region policy optimization for active object recognition, IEEE Transactions on Neural Networks and Learning Systems, 2018

主动目标定位

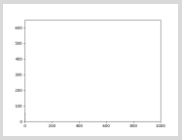
- Active object discovery and localization using sound-induced attention, IEEE Transactions on Industrial Informatics, 2021

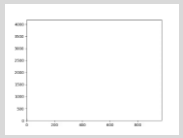
主动触觉识别

- Material identification using tactile perception: A semantics-regularized dictionary learning method, IEEE/ASME Transactions on Mechatronics, 2018

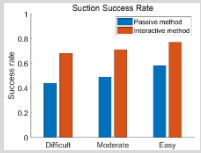
视觉引导的主动触觉感知

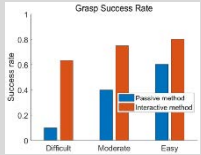






- Visual affordance guided tactile material recognition for waste recycling, IEEE Transactions on Automation Science and Engineering, 2022

主动交互感知

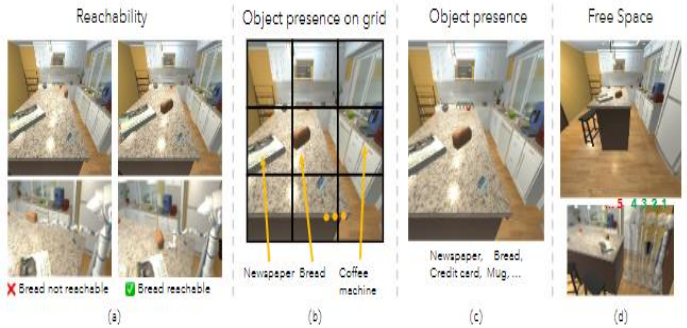


- An interactive perception method for warehouse automation in smart cities, IEEE Transactions on Industrial Informatics, 2021

2 视觉语义导航

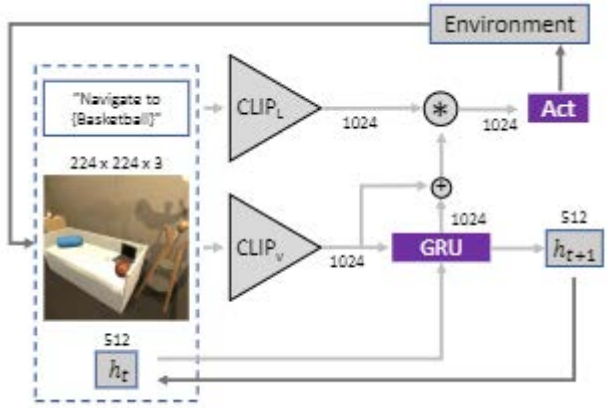
➢ EmbClip

- 在未知场景中进行目标导航时，良好的视觉表示可以实现对导航信息有效的编码，这对于导航非常重要
- 基于CLIP的视觉表示在物体存在、物体存在位置、物体可达性和自由空间等语义上实现了良好的编码



模型结构

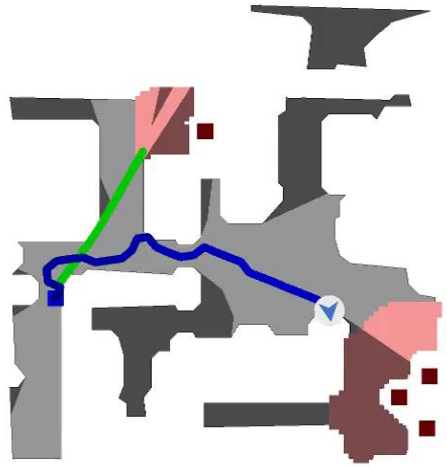
模型将当前观测图像和导航语义分别经过CLIP Vison编码器和 CLIP Language编码器获取特征，特征融合后放入动作选择模型（Act）获取导航动作，另外模型使用GRU处理序列信息

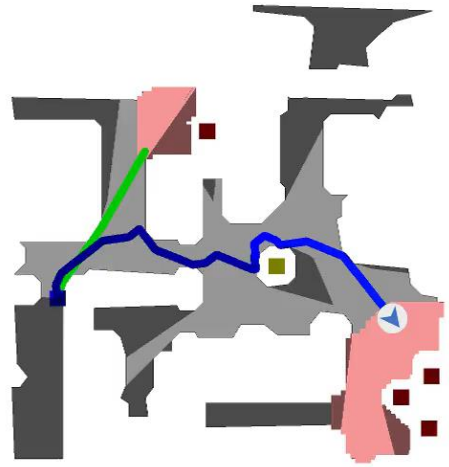


- Simple but effective: Clip embeddings for embodied ai（CVPR 2022）Khandelwal A, Weihs L, Mottaghi R, et al.

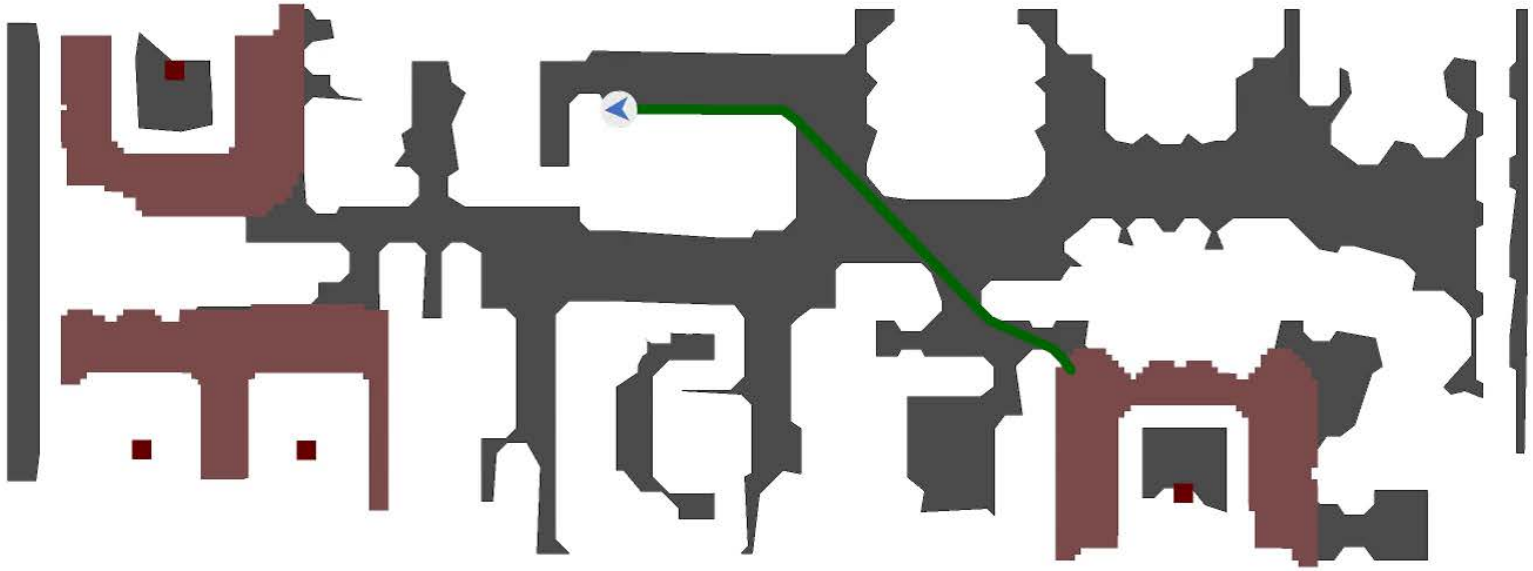
3 具身场景描述

- 脆弱性



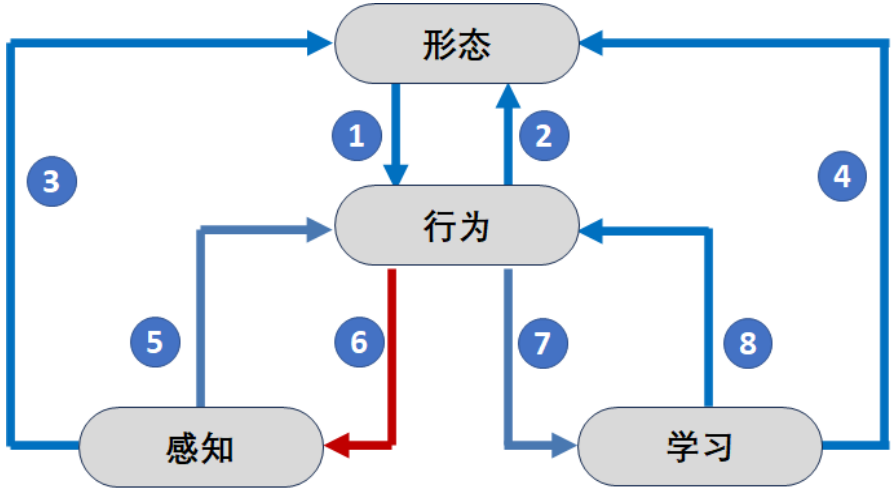


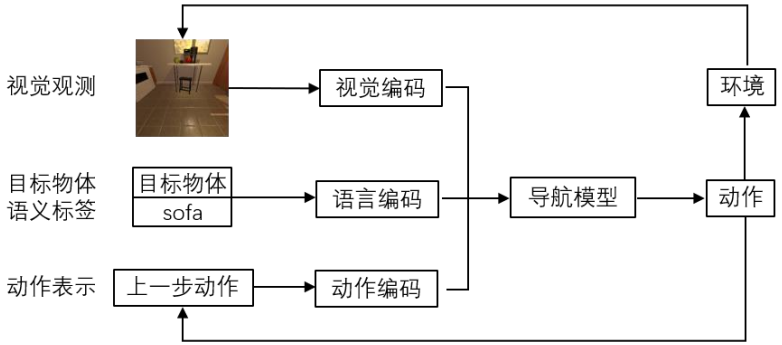

distance_to_goal: 8.05
success: 0.00
spl: 0.00
soft_spl: 0.00
distance_to_goal_reward: 0.00
goal_category: bed

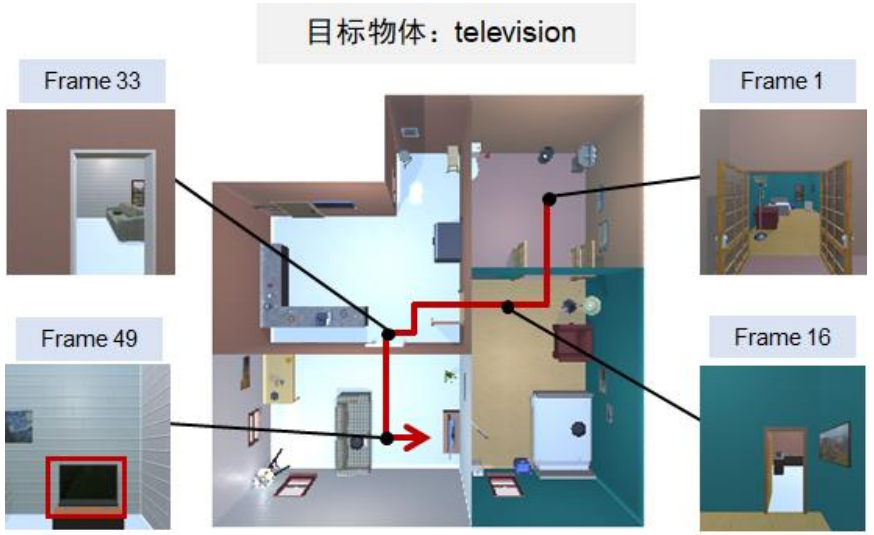


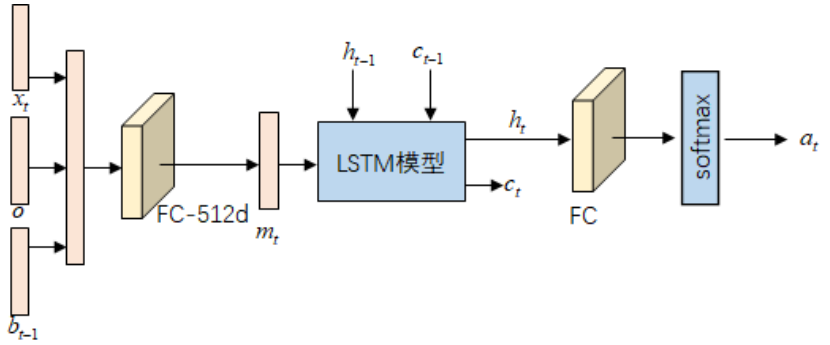
2 视觉语义导航

➢ 小结









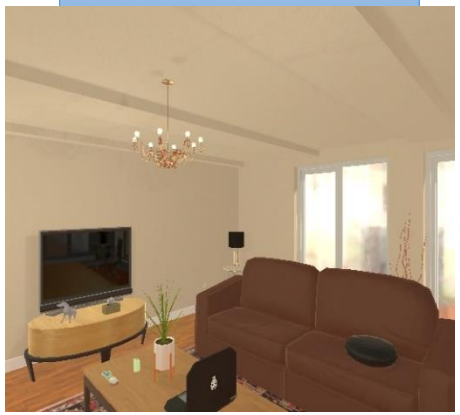
- 场景图片、知识图谱，……

- 从视觉导航到主动感知
- 视觉语义导航
- 具身场景描述
- 具身语言问答

3 具身场景描述

➢ 图像语言描述

Image Captioning

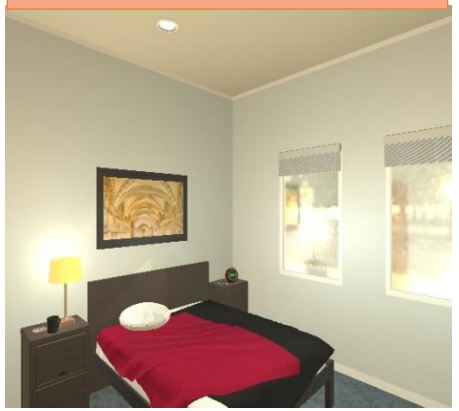


A living room with a couch and a tv.

Dense Captioning

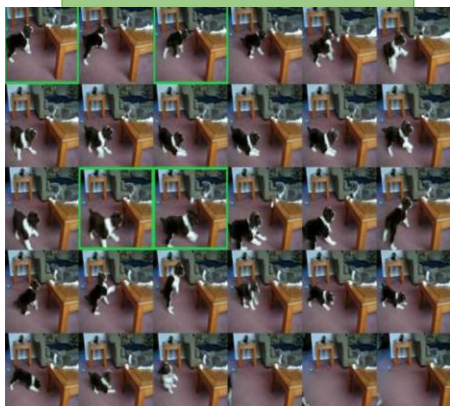
a picture on the wall. the bed is white. the lights are on. pink top on the cake. a small plant in a vase. the chairs are on the table. a white shade. a white chair. the lights are on. a white door.

Image Paragraphing



A bedroom has a large bed inside it and a large plastic bed on which. a large high window is hanging over the bed bed. the bed is round and has a flower on top of it. there is a small lamp lamp on the table that has a picture hanging on it

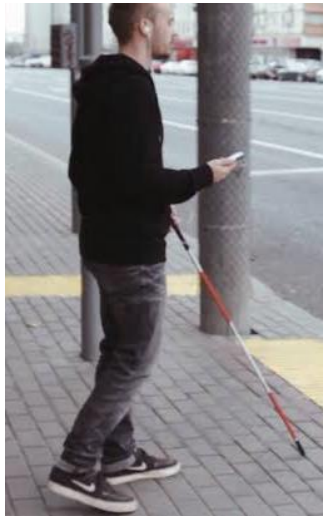
Video Captioning



A cat is playing with a dog.

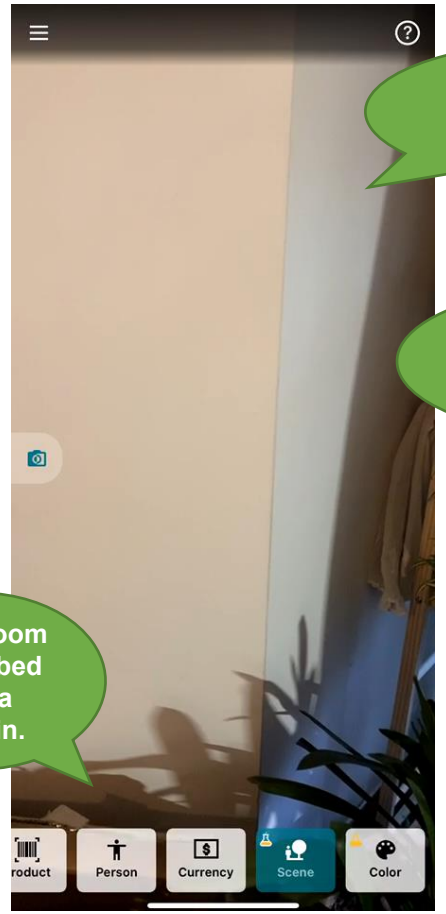
3 具身场景描述

- 图像语言描述



3 具身场景描述

➢ 图像语言描述

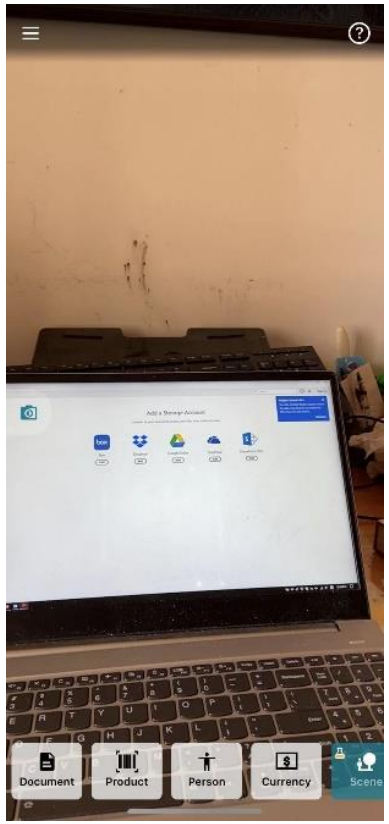


Probably a shower curtain.

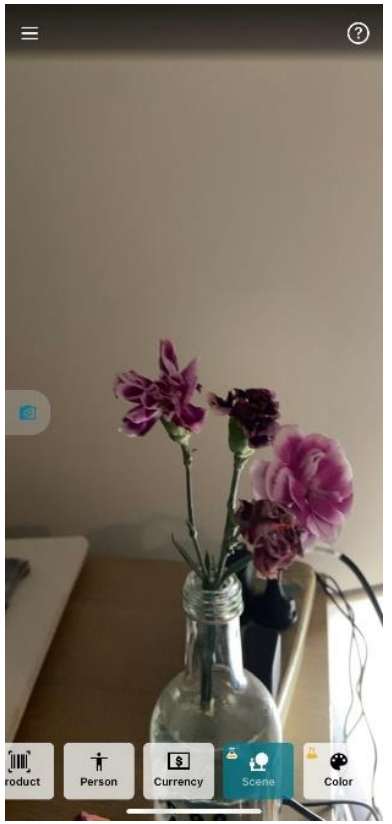
Poor viewpoint detected. Please turn left.

A bedroom with a bed and a curtain.

Probably a close up of a coffee mug.



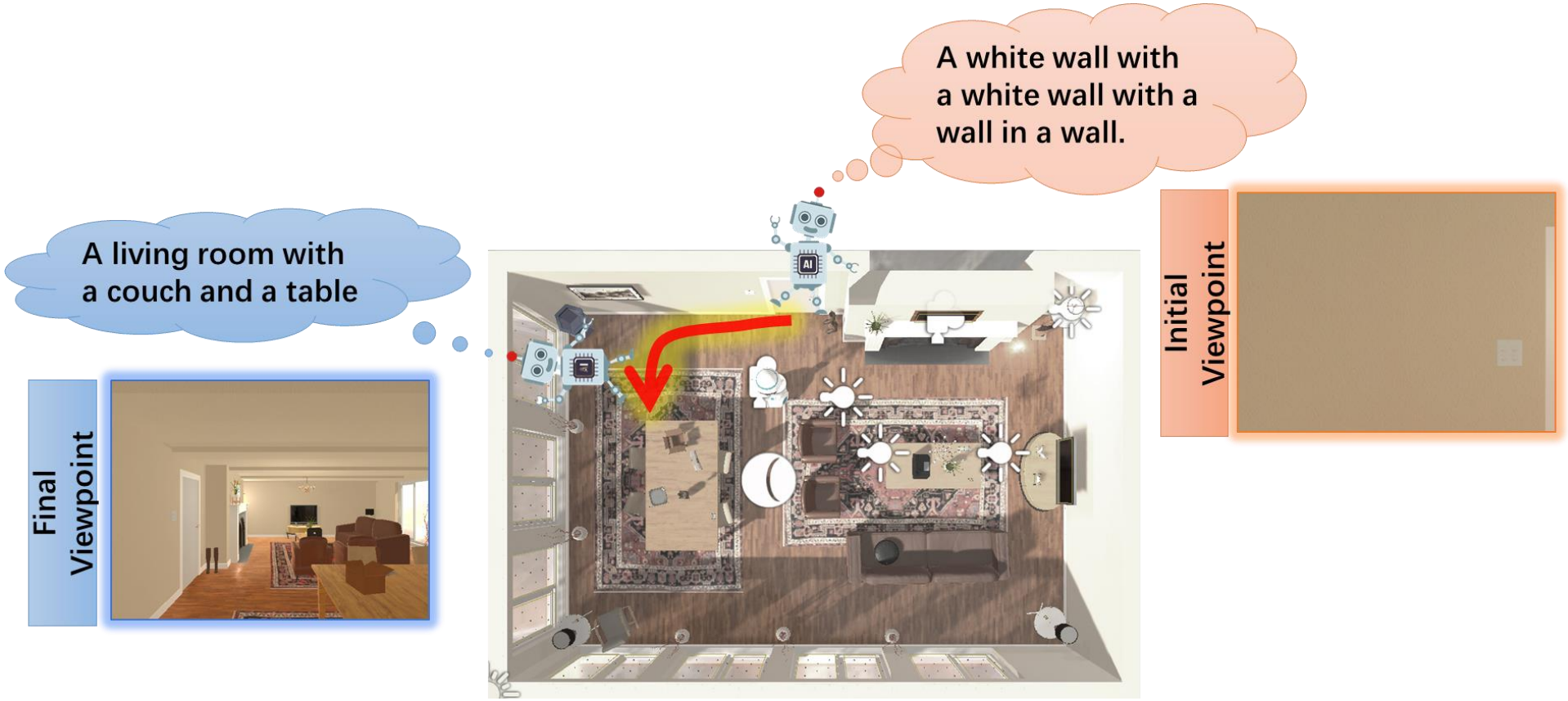
An open laptop sitting on the table.



A vase of flowers on a table.

3 具身场景描述

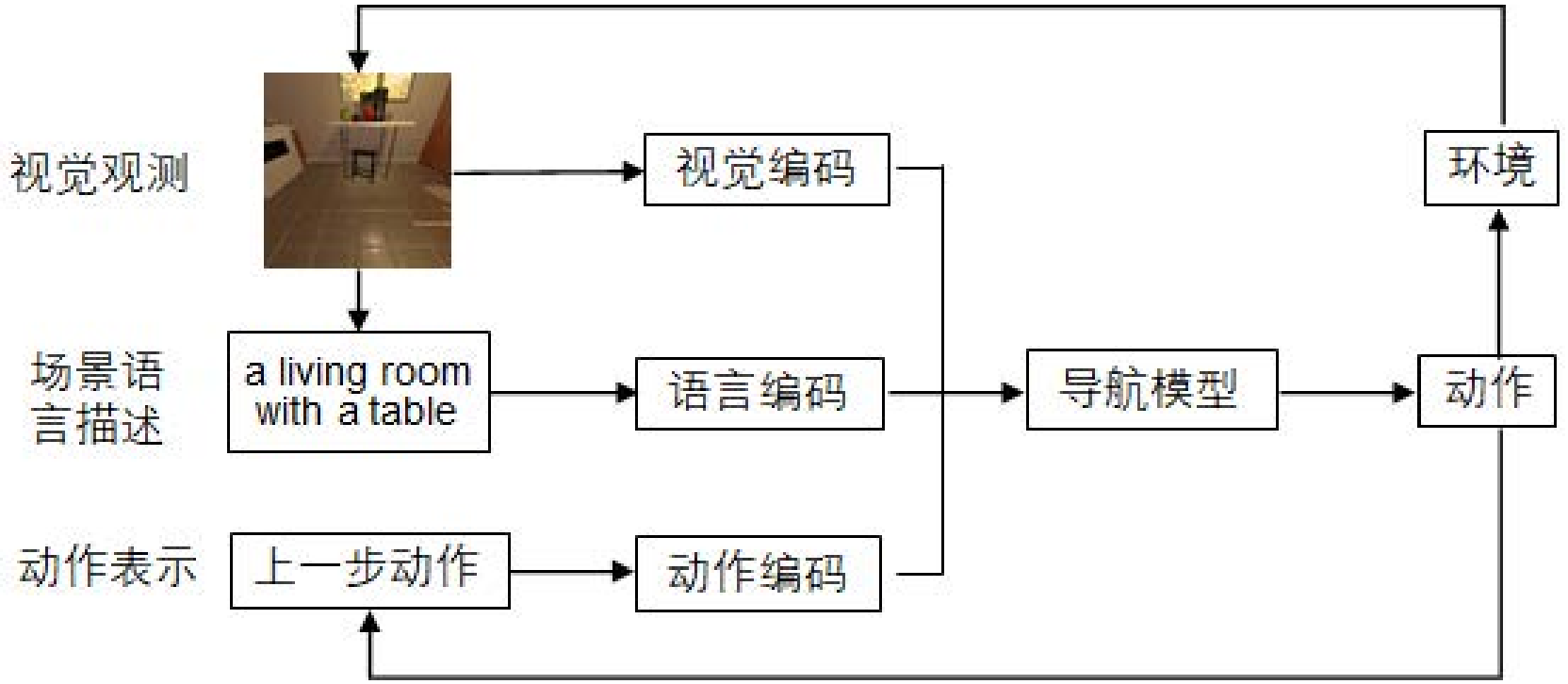
➢ 图像语言描述



- Embodied scene description, RSS, 2020

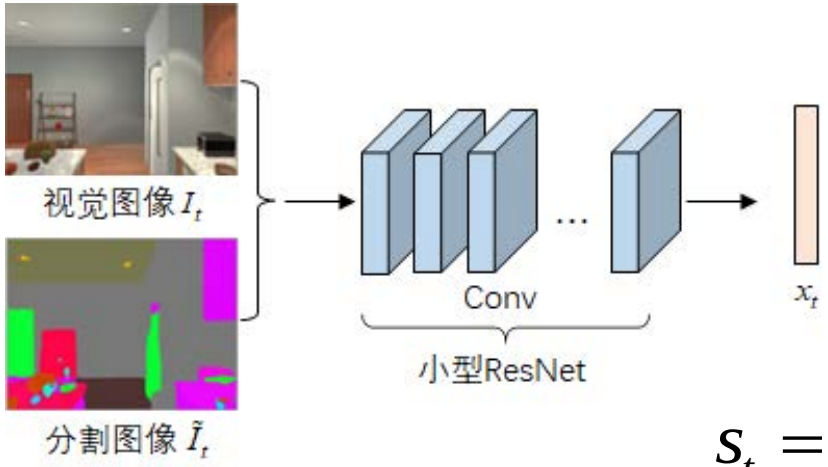
3 具身场景描述

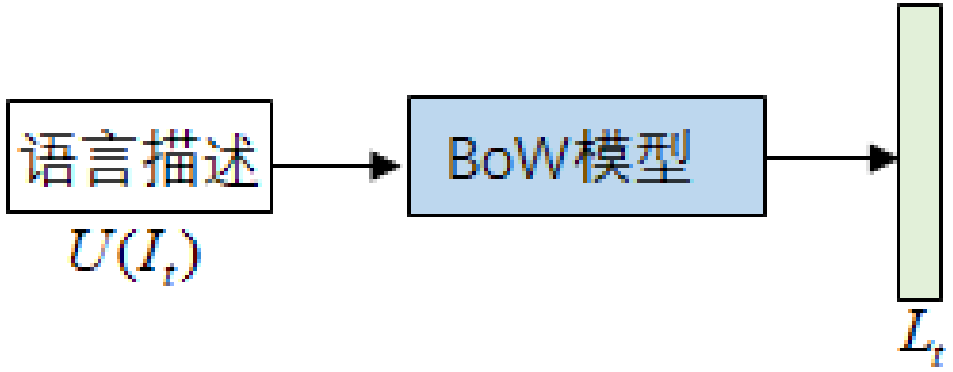
➢ 算法框架



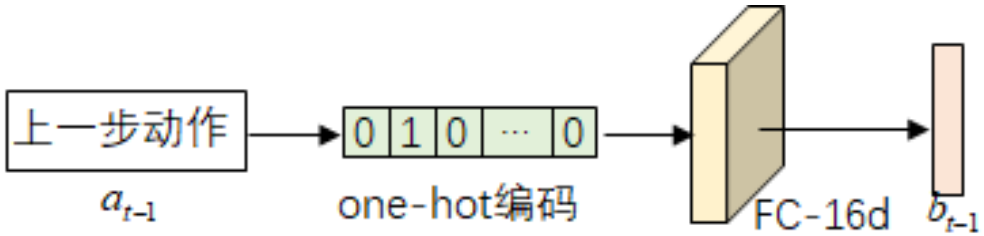
3 具身场景描述

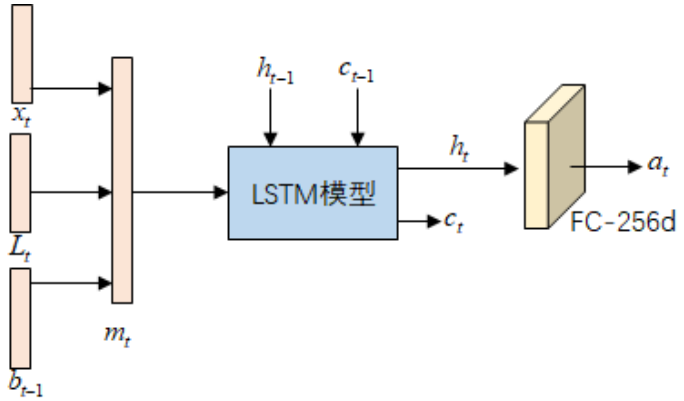
➢ 编码





$$s_t = [ResNet(I_t, \tilde{I}_t); W_L L_t]$$

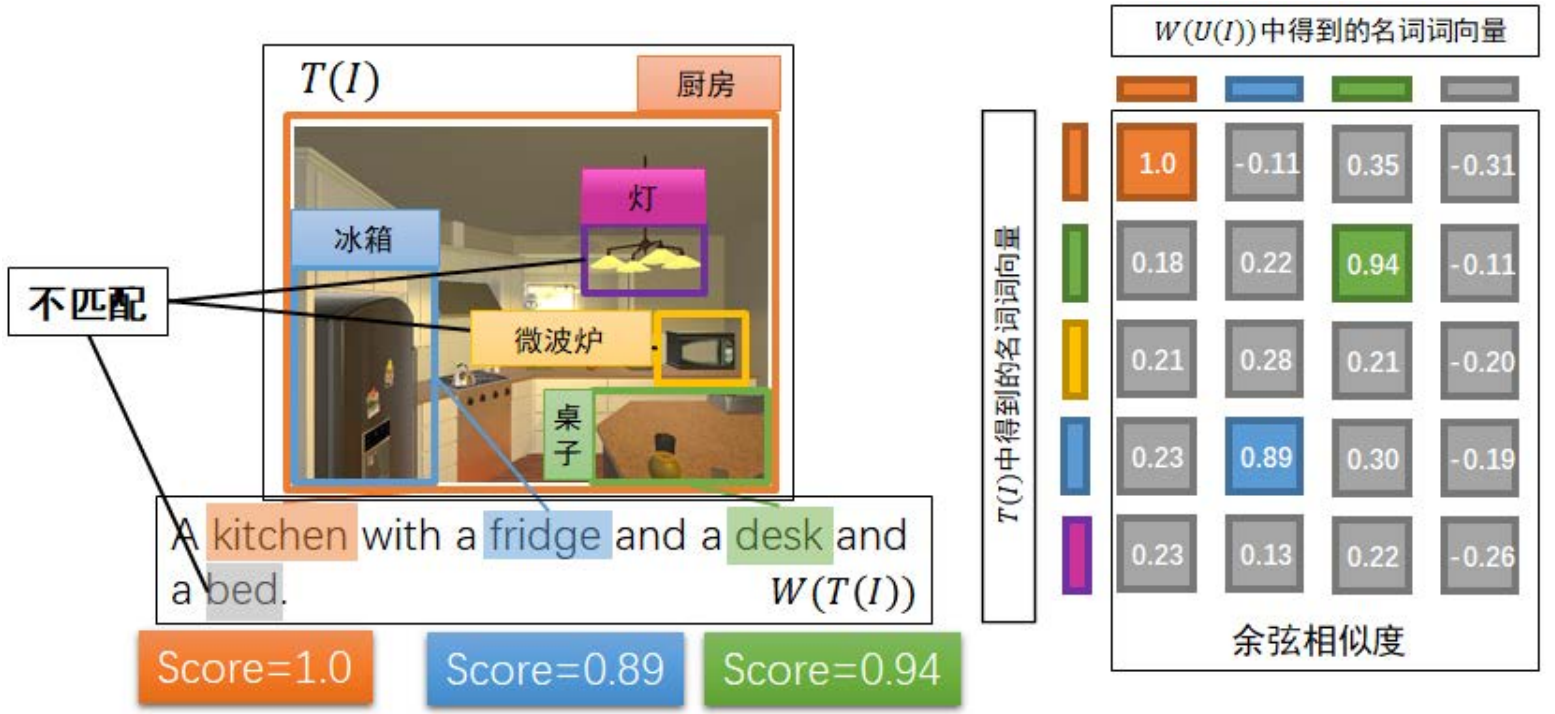




3 具身场景描述

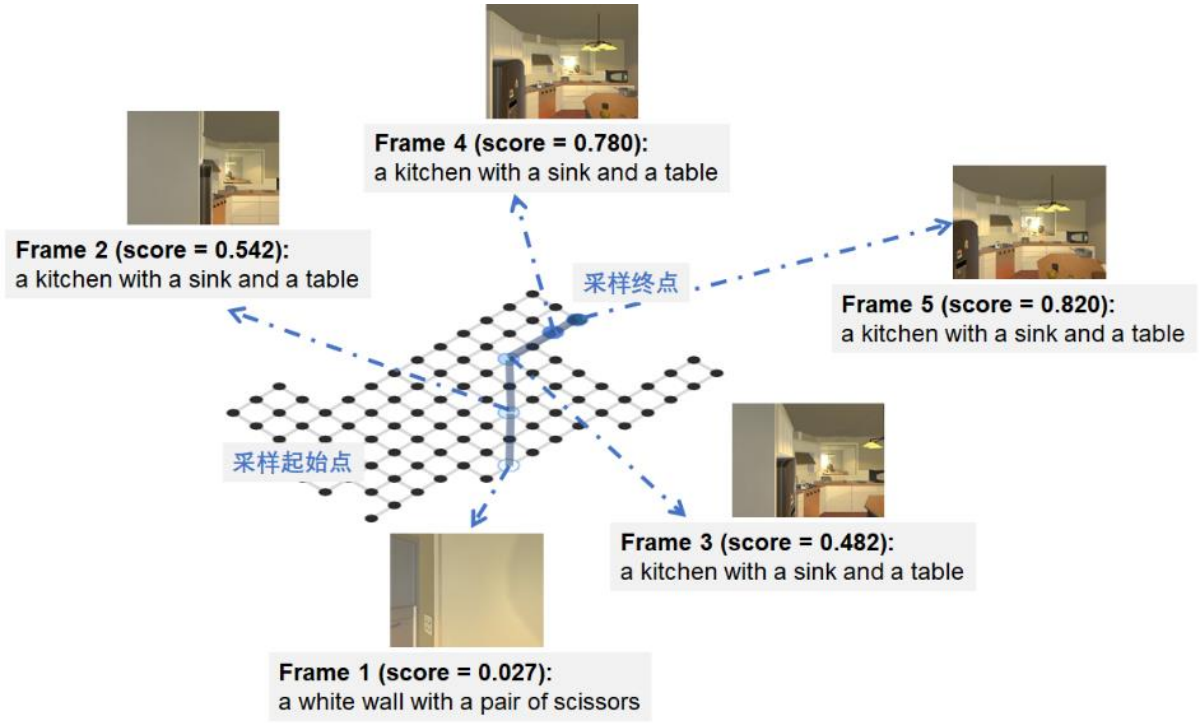
➢ 语言描述的评价

$$score(I) = sim(I, U(I)) + \lambda \frac{|T(I)|}{N}$$



3 具身场景描述

➢ 模型训练：模仿学习+强化学习



$$L_\theta = \sum_{t=1}^{T} -\log \pi_\theta \left(a_t \mid s_1, a_1, \cdots, s_{t-1}, a_{t-1} \right)$$

3 具身场景描述

➢ 实验验证：数据集

将智能体所处环境进行网格化。动作空间由平移运动和旋转运动构成。

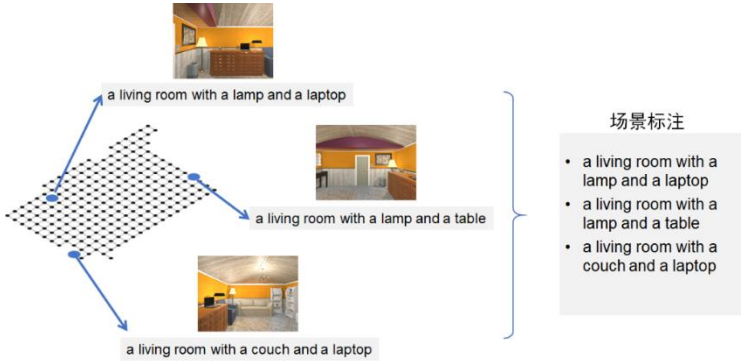
平移运动：包括前、后、左、右、左前、左后、右前、右后，以及停止。每一次平移运动的移动距离固定为网格的边长。

旋转运动：以等间隔进行旋转运动，每一次旋转角度为45°。

3 具身场景描述

➢ 实验验证：数据集

- Number of Steps (NoS)：智能体停止之前走的步数。
- Score of the Last Image (SoL)：触发 Stop 动作的位置的得分。
- Natural Language Metrics：利用为每个场景自动生成的自然语言标注，自然语言任务的指标就可以用来评估所提出的任务。我们选择了几个指标，包括BLEU（Bilingual Evaluation understudy），Meteor，以及ROUGE_L。 其中，BLEU和ROUGE_L指标都是基于n-gram判断生成语句与参考语句相似度的指标，BLEU指标侧重于准确率，ROUGE_L指标侧重于召回率，而Meteor指标同时考虑了准确率与召回率。



$$bleu_n = \frac{\sum_{c \in candidates} \sum_{n-gram \in c} Count_{clip}(n-gram)}{\sum_{c' \in candidates} \sum_{n-gram' \in c'} Count(n-gram')}$$

candidate: the cat sat on the mat

reference: the cat is on the mat

$$bleu_1 = \frac{5}{6} = 0.83$$
$$bleu_2 = \frac{3}{5} = 0.60$$
$$bleu_3 = \frac{1}{4} = 0.25$$
$$bleu_4 = \frac{0}{3} = 0.00$$

- Embodied scene description, RSS, 2020

3 具身场景描述

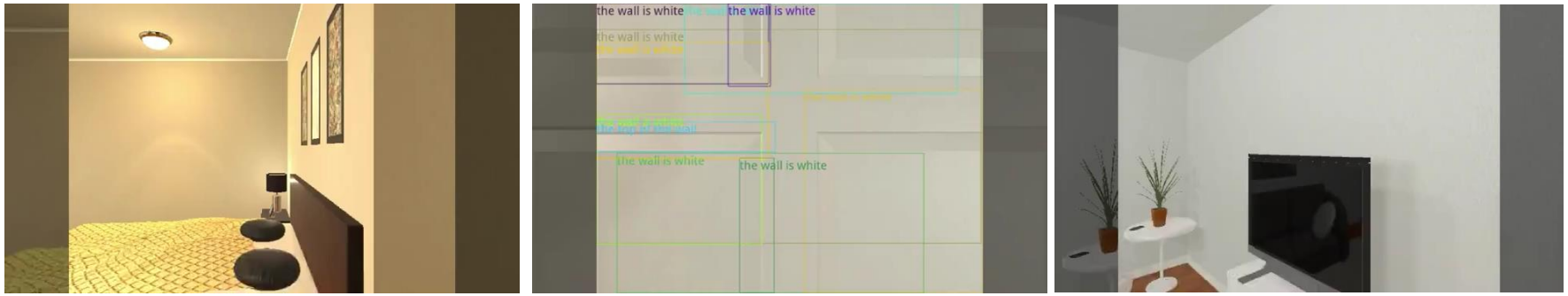
➢ 实验验证：方法比较

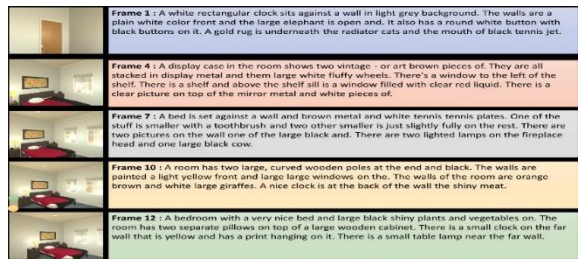
- 随机探索模型：智能体每一步从其动作空间中随机选择探索动作。
- IL (RGB)：利用模仿学习和 RGB 图像训练动作生成模型。
- IL (RGB+Segm.)：利用模仿学习同时使用 RGB 图像和语义分割图训练动作生成模型。
- IL+RL (RGB+Segm.)：同时以RGB图像和分割图像作为输入，使用模仿学习训练和强化学习微调模型的范式训练动作生成模型。

	NoS(↓)	SoL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L
Random	26.97	0.44	0.57	0.46	0.36	0.26	0.29	0.61
IL(RGB)	38.75	0.79	0.75	0.66	0.56	0.42	0.46	0.79
IL(RGB+Segm)	16.99	0.81	0.78	0.69	0.59	0.43	0.48	0.82
IL+RL(RGB+Segm)	15.27	0.81	0.77	0.69	0.58	0.43	0.48	0.82

3 具身场景描述

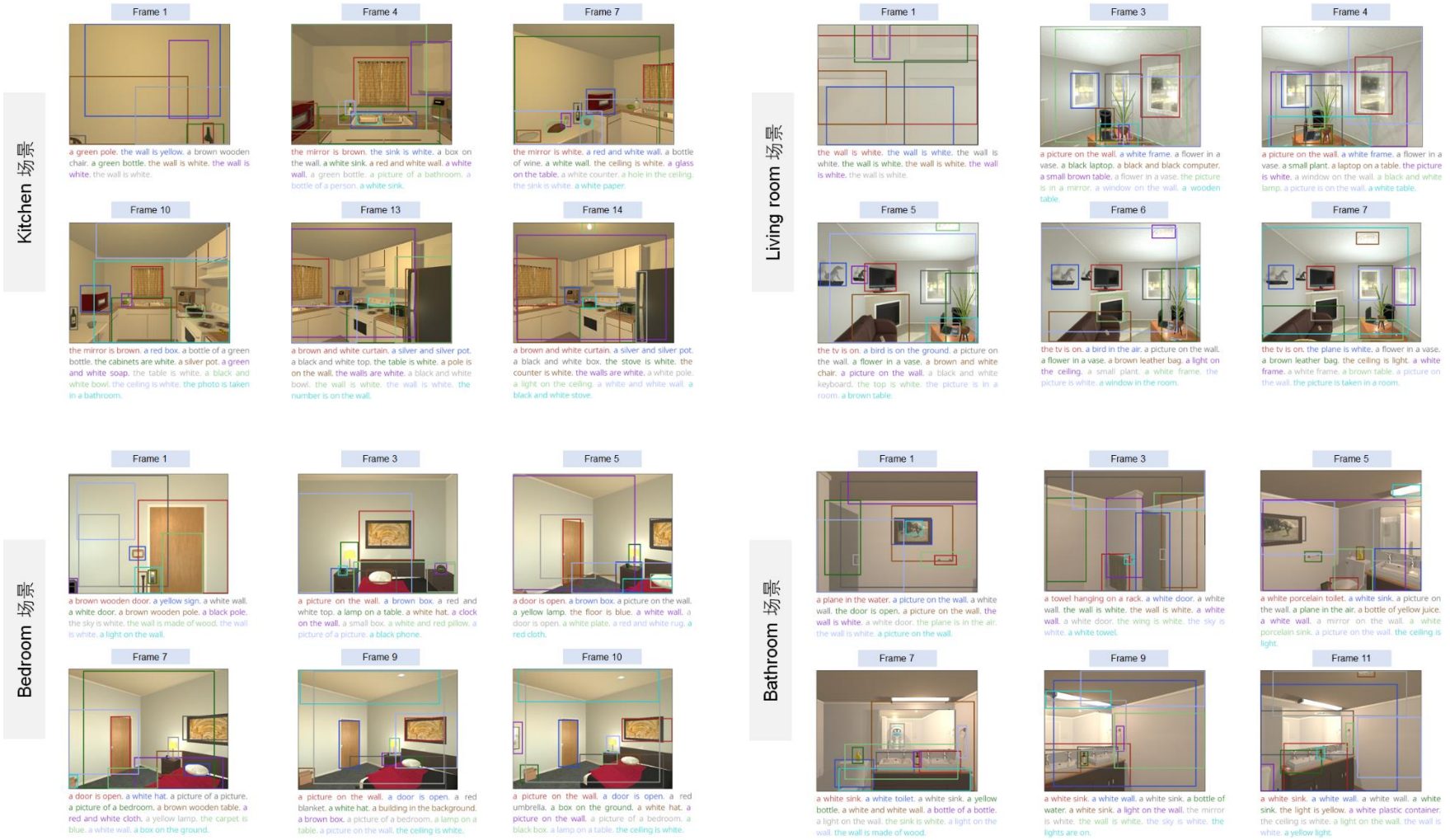
➢ 实验验证





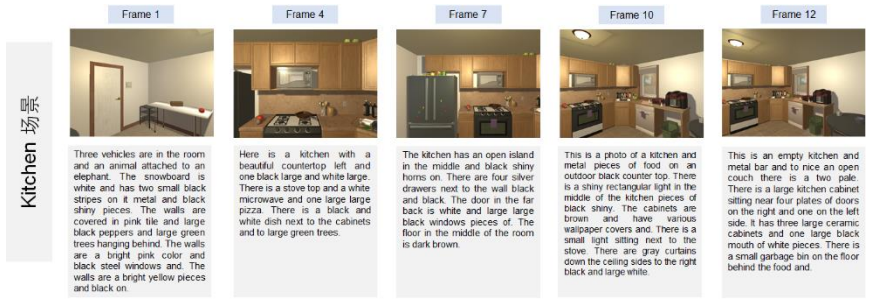
3 具身场景描述

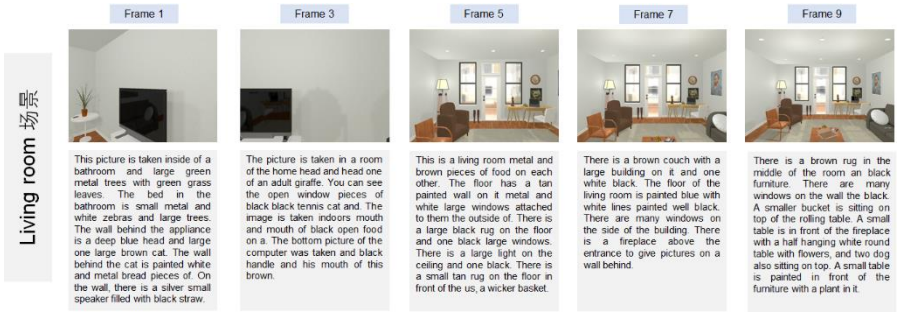
➢ 实验验证：Dense Captioning

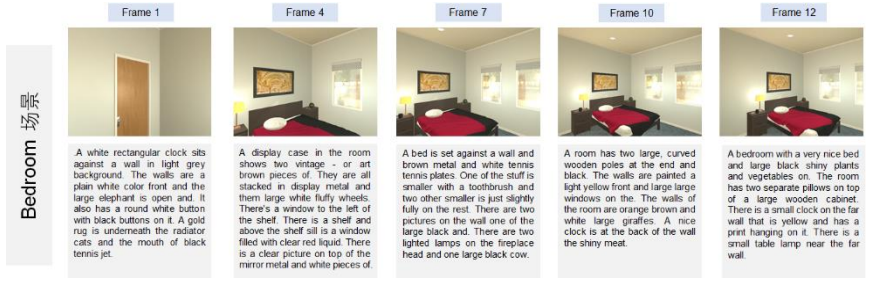


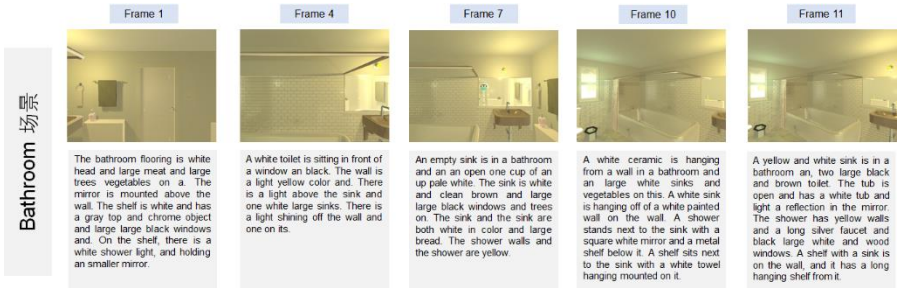
3 具身场景描述

➢ 实验验证：Image Paragraphing





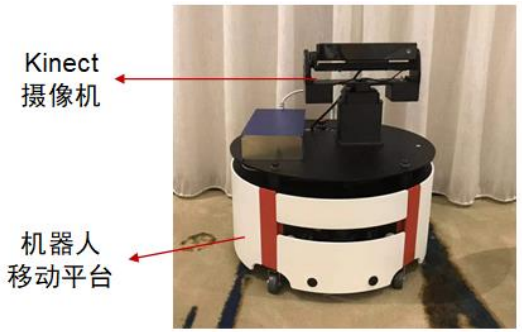




	Nos	SoL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L
Dense captioning	14.41	0.706 3	0.7637	0.6315	0.5201	0.4275	0.2875	0.4815
Image paragraphing	13.28	0.693 6	0.5600	0.3568	0.2224	0.1510	0.1706	0.3601

3 具身场景描述

➢ 物理验证

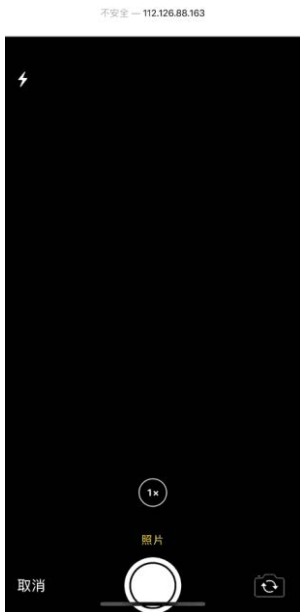


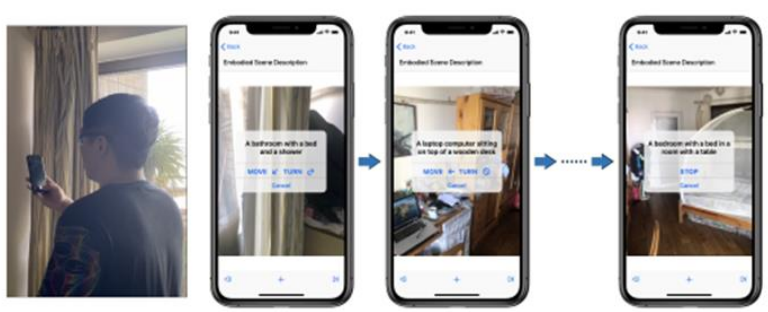
第三视角图像

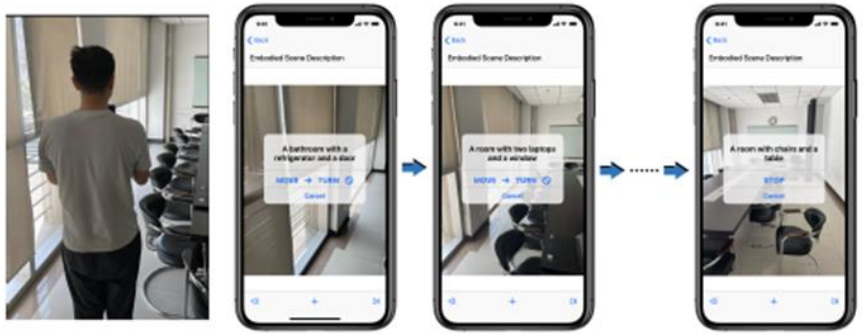
第一视角图像

3 具身场景描述

➢ 交互验证

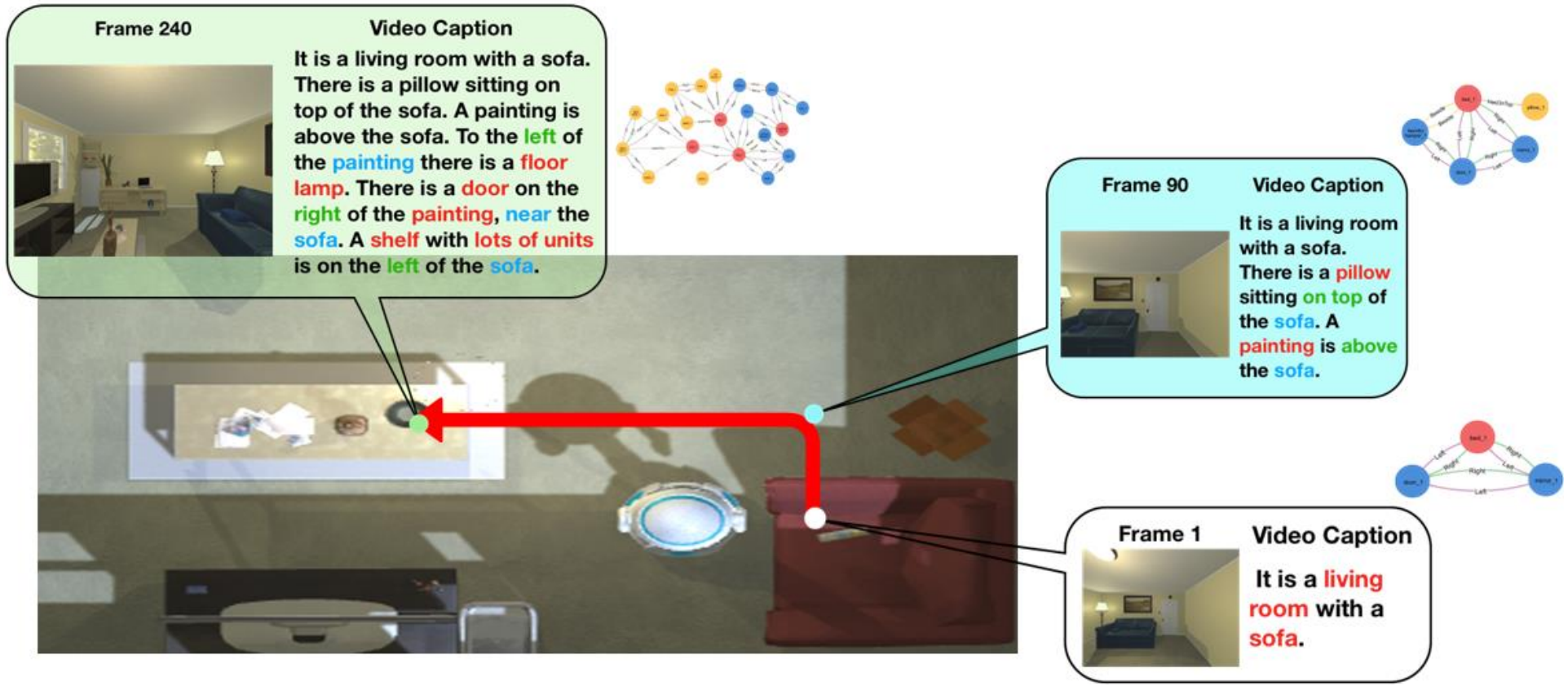






3 具身场景描述

➢ 拓展：视频场景描述



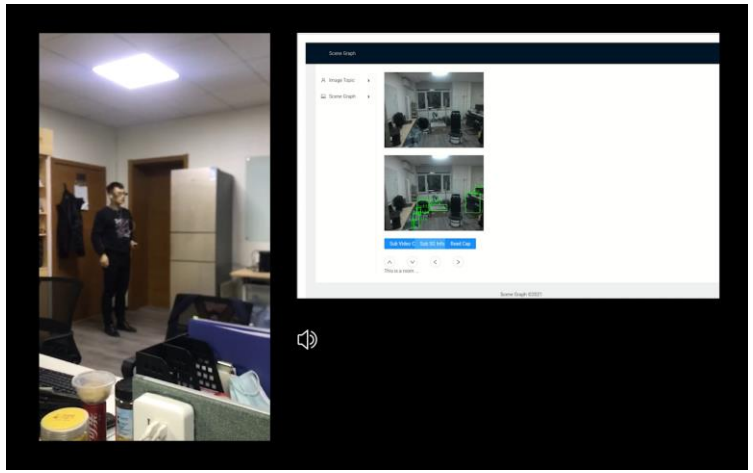
- Robotic indoor scene captioning from streaming video, ICRA, 2021

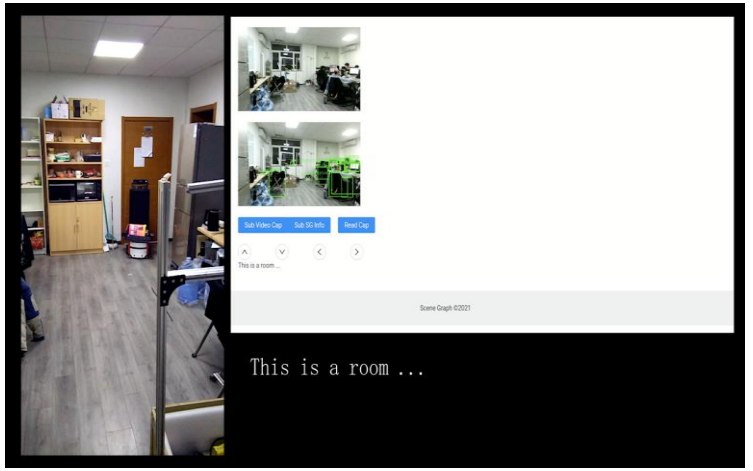
3 具身场景描述

➢ 拓展：视频场景描述

It is a living room with a sofa.There is a small table in front of the sofa.There is a pillow sitting on top of the sofa.There is a painting above the sofa.To the left of the painting there is a floor lamp .

It is a living room with two chairs.There is also a laptop in the room .





3 具身场景描述

➢ 拓展：红外场景描述

A blurry photo of a person in a rain.

A city street with cars and a traffic light.

A man riding a snowboard down a snow covered slope.

A man and a woman are sitting on a bench.

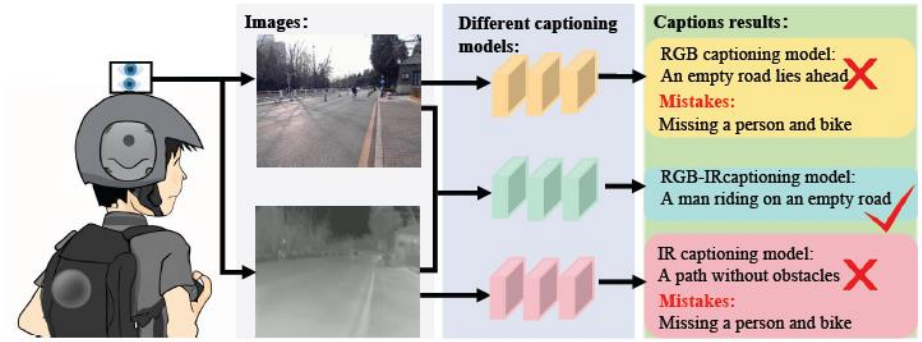
Outdoor Scene

Indoor Scene

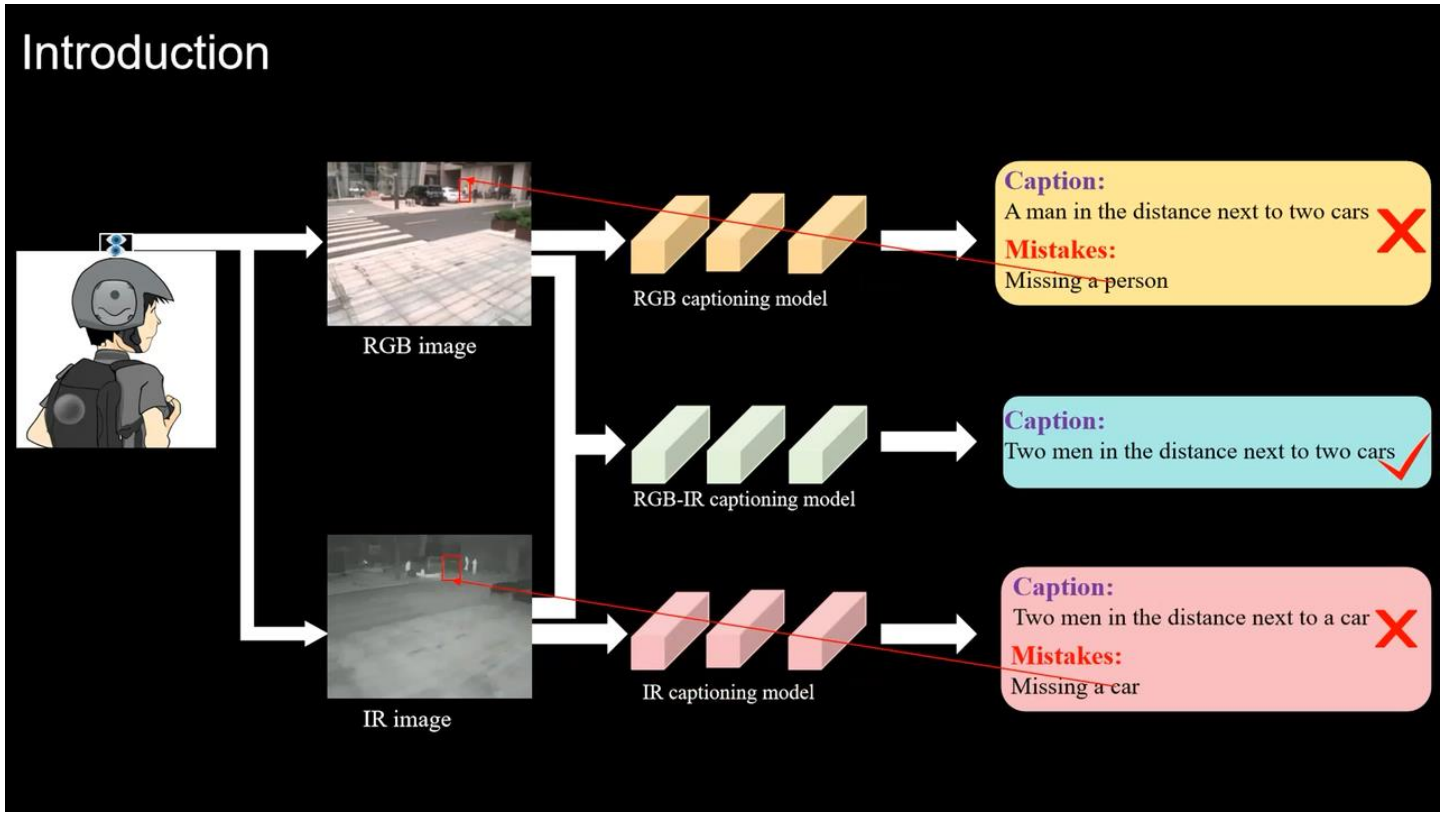
- Infrared image captioning with wearable device, ICRA, 2023

3 具身场景描述

➢ 多模态


Images:
Different captioning models:
Captions results:
RGB captioning model:
An empty road lies ahead
Mistakes:
Missing a person and bike
RGB-IRcaptioning model:
A man riding on an empty road
IR captioning model:
A path without obstacles
Mistakes:
Missing a person and bike

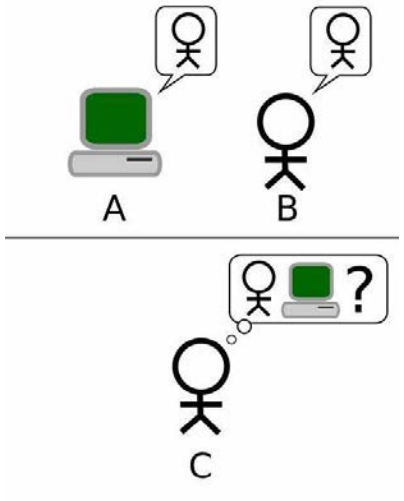

Device list:
① Portable backpacks
② Portable Power
③ Ergonomic helmet
④ Infrared camera
⑤ Visible light camera
⑥ NVIDIA Jetson TX2
⑦ Mini Bluetooth Controller
⑧ Wired headset
Demo

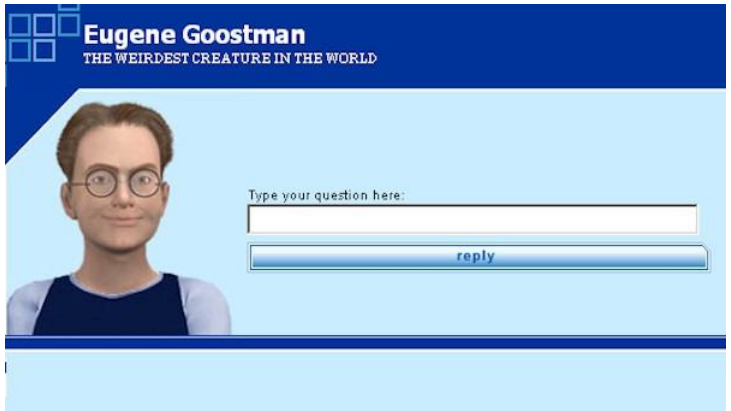

Introduction
RGB image
IR image
RGB captioning model
RGB-IR captioning model
IR captioning model
Caption:
A man in the distance next to two cars
Mistakes:
Missing a person
Caption:
Two men in the distance next to two cars
Caption:
Two men in the distance next to a car
Mistakes:
Missing a car

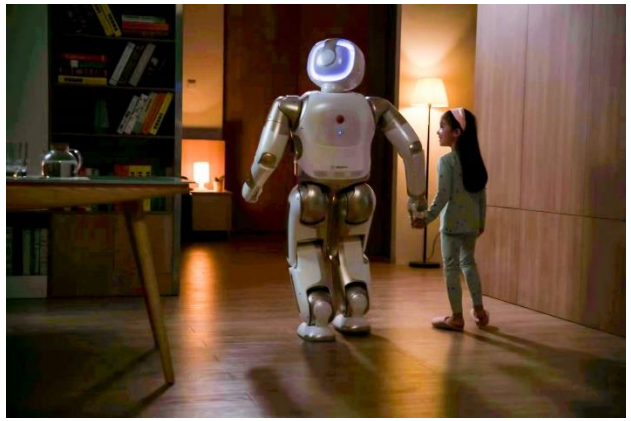
- 从视觉导航到主动感知
- 视觉语义导航
- 具身场景描述
- 具身语言问答

4 具身语言问答

➢ 图灵测试







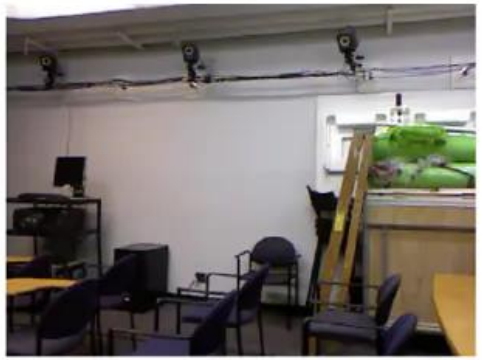
4 具身语言问答

➢ Visual Question Answering

Visual Question Answer (VQA) 是对视觉图像的自然语言问答，作为视觉理解 (Visual Understanding) 的一个研究方向，连接着视觉和语言，模型需要在理解图像的基础上，根据具体的问题然后做出回答。

Q: How many white objects in this picture ?

A: 9



Q: What color is the chair in front of the wall on the left side of the stacked chairs ?

A: blue

Q: What is the largest white object on the left side of the picture ?

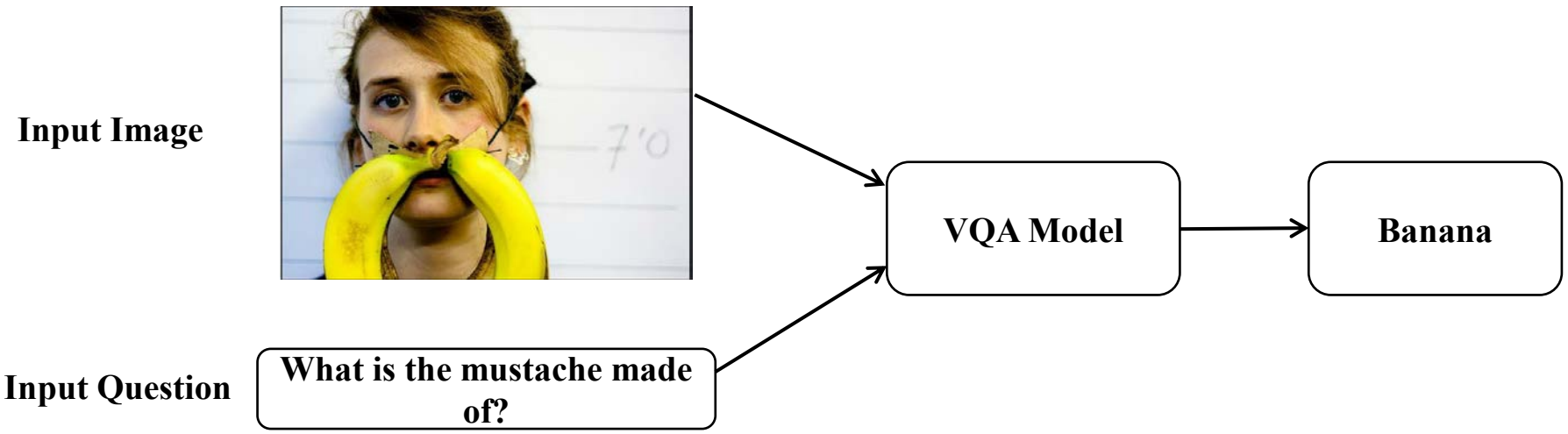
A: printer

4 具身语言问答

➢ Visual Question Answering

VQA Visual Question Answering

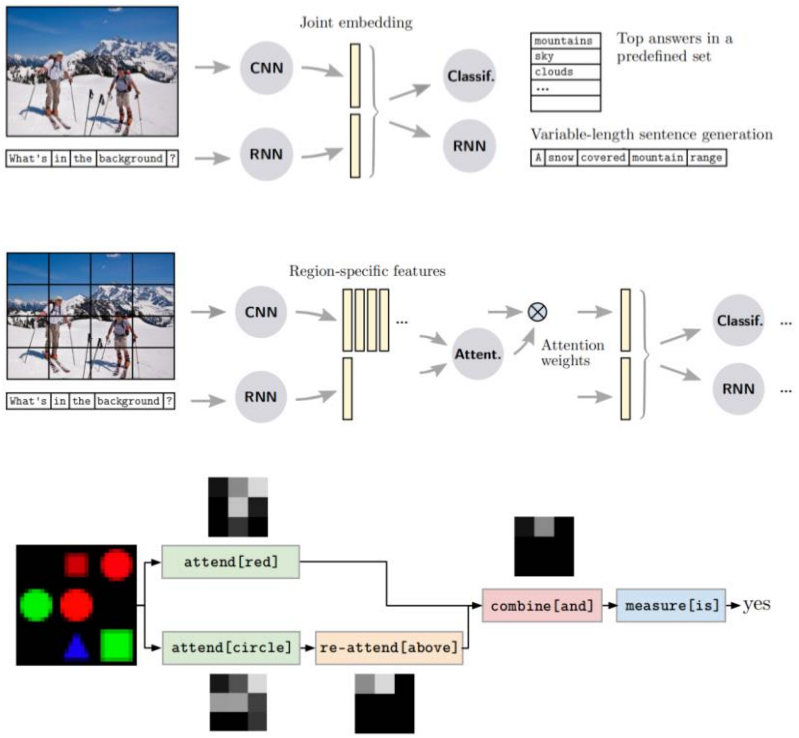
https://visualqa.org/



一个VQA系统以一张图片和一个关于这张图片形式自由、开放式的自然语言问题作为输入，以生成的自然语言答案作为输出。

4 具身语言问答

➢ Visual Question Answering



- **Malinowski等人作为提出“开放世界”视觉问答的第一个尝试人之一，在[1]中描述了一种在贝叶斯公式中结合语义文本解析和图像分割的方法**
- **Tu等人在[2]中对VQA的另一个早期尝试是基于文本和视频的联合解析图。**
- **在[3]中，German等人提出了一种自动的“查询生成器”，它对带注释的图像进行训练，然后从任何给定的测试图像中生成一系列二值问题。**

主要方法：

- **联合嵌入方法: Joint embedding approaches**
- **注意力机制: Attention mechanisms**
- **组合模型: Compositional Models**
- **知识增强方法: Models using external knowledge base**

- **[1] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Proc.Advances in Neural Inf. Process. Syst., pages 1682–1690, 2014.**
- **[2] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queries. IEEE Trans. Multimedia, 21(2):42–70, 2014.**
- **[3] D. Geman, S. Geman, N. Hallonquist, and L. Y ounes. Visual Turing test for computer vision systems. Proceedings of the National Academy of Sciences, 112(12):3618–3623, 2015.**

4 具身语言问答

- **QA(ChatGPT)+Robot**

4 具身语言问答

➢ 问题描述

Embodied question answering（EQA）在3D环境中，智能体出现在随机位置，然后问智能体一个问题（汽车是什么颜色的？）为了找到答案，智能体首先探索环境，以第一人称视角收集视觉信息，然后回答问题（橙色）

Input：
First-person view visual Scene (no GPS, no map)
question

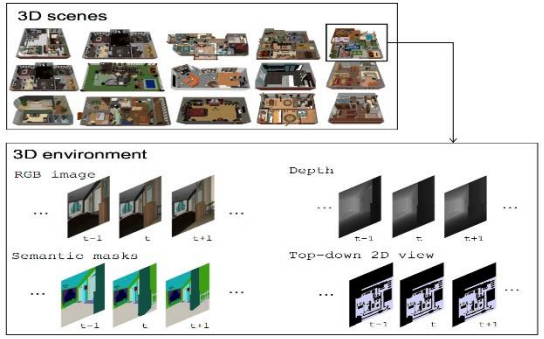
What color is the car?

Output:
Actions: forward, forward, turn left …… stop
Answer : orange

- Embodied question answering. CVPR, 2018

4 具身语言问答

➢ 数据集



(b)House3D environment

(a) SUNCG dataset

◈containing more than 40K manually created indoor environments.

◈ All these scenes are composed of individually labeled 3D objects.

◈ allowing us to compute full volumetric ground truth labels

4 具身语言问答

➢ 数据集

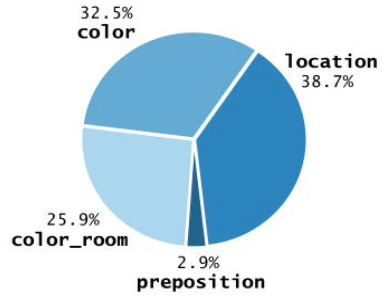
EQA v1 dataset：

●consists of 4 question types– location, color, color_room, preposition.

● consists of over 5000 question across over 750 environments

● 7 unique room types，45 unique objects

● Approximately 6 questions are asked per environment on average, 22 at most, and 1 at fewest.



上图显示了数据集分割和问题类型分布

EQA v1

location:	*'What room is the <OBJ> located in?'*
color:	*'What color is the <OBJ>?'*
color_room:	*'What color is the <OBJ> in the <ROOM>?'*
preposition:	*'What is <on/above/below/next-to> the <OBJ> in the <ROOM>?'*
existence:	*'Is there a <OBJ> in the <ROOM>?'*
logical:	*'Is there a(n) <OBJ1> and a(n) <OBJ2> in the <ROOM>?'*
count:	*'How many <OBJs> in the <ROOM>?'*
room_count:	*'How many <ROOMs> in the house?'*
distance:	*'Is the <OBJ1> closer to the <OBJ2> than to the <OBJ3> in the <ROOM>?'*

gym	dining room
patio	living room
office	bathroom
lobby	bedroom
garage	elevator
kitchen	balcony

(b) Queryable Rooms

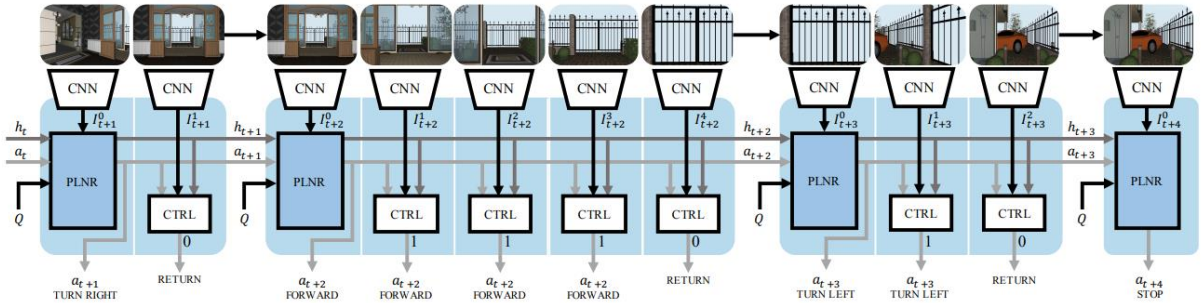
rug	piano	dryer	computer	fireplace	whiteboard	bookshelf	wardrobe cabinet
pan	toilet	plates	ottoman	fish tank	dishwasher	microwave	water dispenser
bed	table	mirror	tv stand	stereo set	chessboard	playstation	vacuum cleaner
cup	xbox	heater	bathtub	shoe rack	range oven	refrigerator	coffee machine
sink	sofa	kettle	dresser	knife rack	towel rack	loudspeaker	utensil holder
desk	vase	shower	washer	fruit bowl	television	dressing table	cutting board
ironing board	food processor						

(c) Queryable Objects

- Embodied question answering. CVPR, 2018

4 具身语言问答

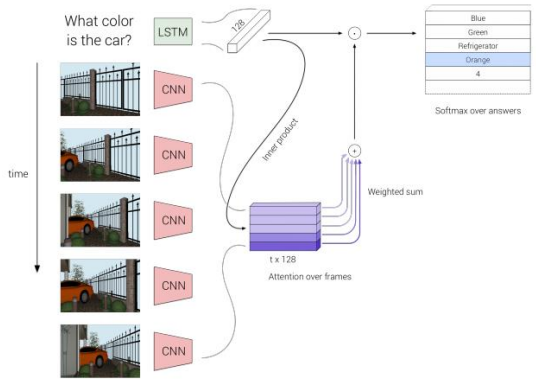
➢ 基本方法



问题回答模块以导航帧和问题为条件，计算最后五帧的点积注意力，并将图像特征的注意力加权组合与问题编码相结合来预测答案

导航模块PACMAN导航器将导航分解为规划器和控制器。
计划者选择动作，控制器执行这些动作的次数不定。这使规划者可以在更短的时间范围内进行操作，从而增强了梯度流。使用第一人称时间的RGB和depth图像作为输入，输出一系列的动作：

$$< a_{t+1}, a_{t+2}, \cdots\cdots \text{stop} >$$



4 具身语言问答

➢ 评价标准

T: 步长，即初始位置到目标位置的步数（向前一步最多0.25米)

d_T 在T的情况下导航终点到目标物之间的距离

d_Δ 在T的情况下导航从起始到终止位置朝目标方向移动的距离

d_{min} 在T的情况下结束位置到目标的最小距离

Evaluate Navigation in Embodied QA

为了评估不同难度的导航性能与问答性能指标，原作者利用不同距离目标位置的步长来划分不同的难度等级即（10步， 30步，50步三个难度等级）

MR：准确答案在智能体回答的基于置信度的答案列表中的排名

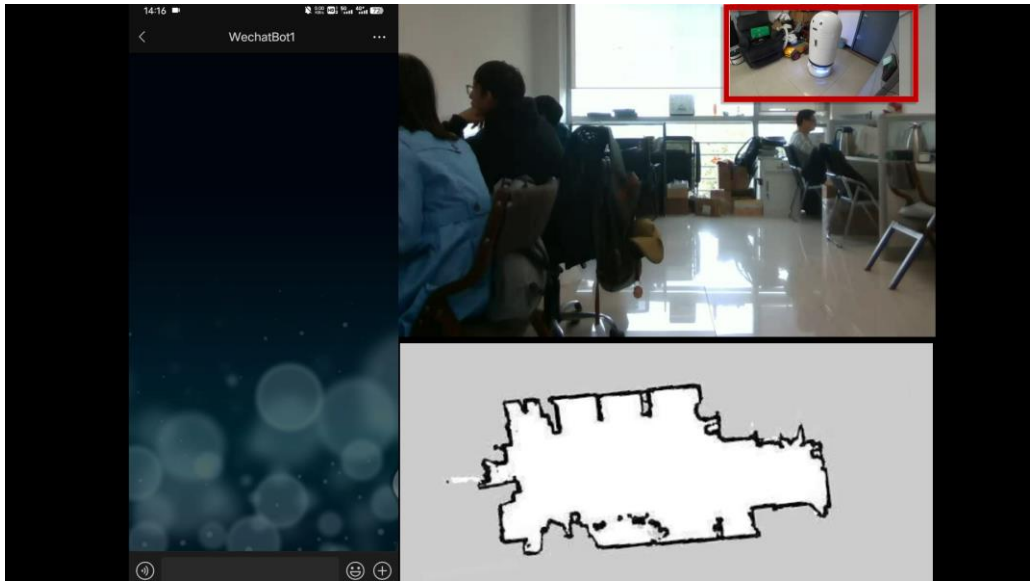
Evaluate QA in Embdoied QA

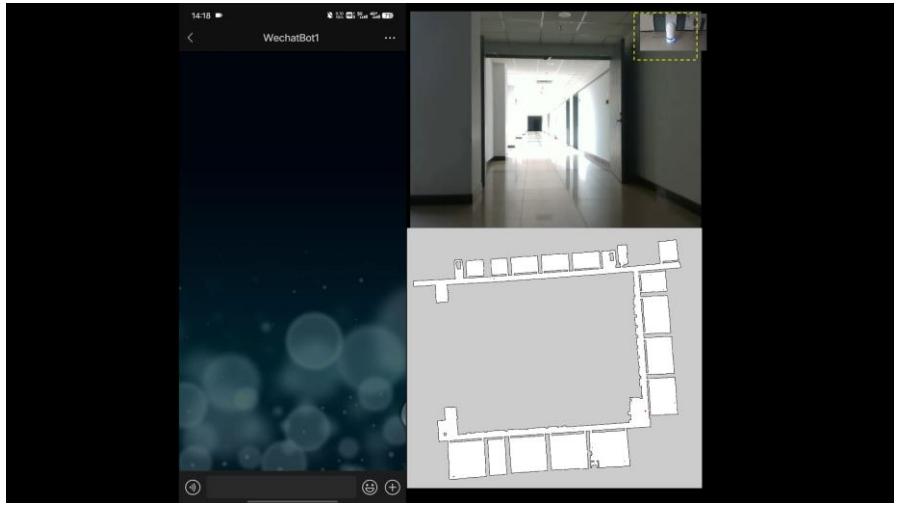
		Navigation																		QA		
		d_T			d_Δ			d_{min}			%r_T			%$r_{\leftarrow}$			%stop			MR		
		T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}
Baselines	Reactive	2.09	2.72	3.14	-1.44	-1.09	-0.31	0.29	1.01	1.82	50%	49%	**47%**	52%	53%	48%	-	-	-	3.18	3.56	3.31
	LSTM	1.75	2.37	2.90	-1.10	-0.74	-0.07	0.34	1.06	2.05	55%	53%	44%	59%	57%	50%	80%	75%	80%	3.35	3.07	3.55
	Reactive+Q	1.58	2.27	2.89	-0.94	-0.63	-0.06	0.31	1.09	1.96	52%	51%	45%	55%	57%	**54%**	-	-	-	3.17	3.54	3.37
	LSTM+Q	1.13	2.23	2.89	-0.48	-0.59	-0.06	0.28	0.97	1.91	**63%**	53%	45%	64%	59%	**54%**	80%	71%	68%	3.11	3.39	3.31
Us	ACT+Q	**0.46**	**1.50**	**2.74**	**0.16**	**0.15**	**0.12**	0.42	1.42	2.63	58%	54%	45%	60%	56%	46%	**100%**	**100%**	**100%**	**3.09**	3.13	3.25
	ACT+Q-RL	1.67	2.19	2.86	-1.05	-0.52	0.01	**0.24**	**0.93**	**1.94**	57%	**56%**	45%	**65%**	**62%**	52%	32%	32%	24%	3.13	**2.99**	**3.22**
Oracle	HumanNav*	0.81	0.81	0.81	0.44	1.62	2.85	0.33	0.33	0.33	86%	86%	86%	87%	89%	89%	-	-	-	-	-	-
	ShortestPath+VQA	-	-	-	0.85	2.78	4.86	-	-	-	-	-	-	-	-	-	-	-	-	3.21	3.21	3.21

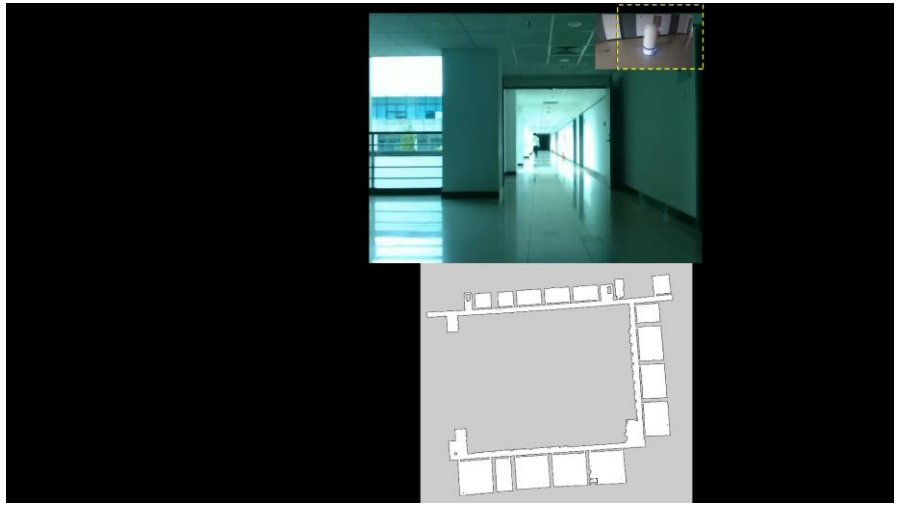
- Embodied question answering. CVPR, 2018

4 具身语言问答

➢ 评价标准







4 具身语言问答

➢ 前沿：层次化导航

Neural Modular Control for Embodied Question Answering

引入了一个用于具体问题回答的分层策略。给定一个问题(“客厅里的沙发是什么颜色的?”)和观察，我们的主策略预测一个子目标序列——出房间，查找房间[生活]，查找对象[沙发]，回答——然后由专门的子策略执行，导航到目标对象并回答问题(“灰色”)。

给定一个问题，智能体的目标是预测一系列导航子目标，并执行它们，最终找到目标对象并给出正确答案。因此提出NMC作为为每个子目标生成子目标和子策略的主策略。每个子目标被分解为一个任务和一个参数 $< g_{subgoal}, g_{arguments} >$。有4个可能的任务-出口房间，寻找房间，寻找对象和回答。

Subgoal	Argument(s)	Description	Success
Exit-room	None	When there is only 1 door in spawn room, or 1 door other than door entered through in an intermediate room; agent is forced to use the remaining door.	Stopping after exiting through the correct door.
Find-room	Room name (*gym, kitchen, ...*)	When there are multiple doors and the agent has to search and pick the door to the target room.	Stopping after entering target room.
Find-object	Object name (*oven, sofa, ...*)	When the agent has to search for a specific object in room.	Stopping within 0.75m of the target object.
Answer	None	When the agent has to provide an answer from the answer space.	Generating the correct answer to the question.

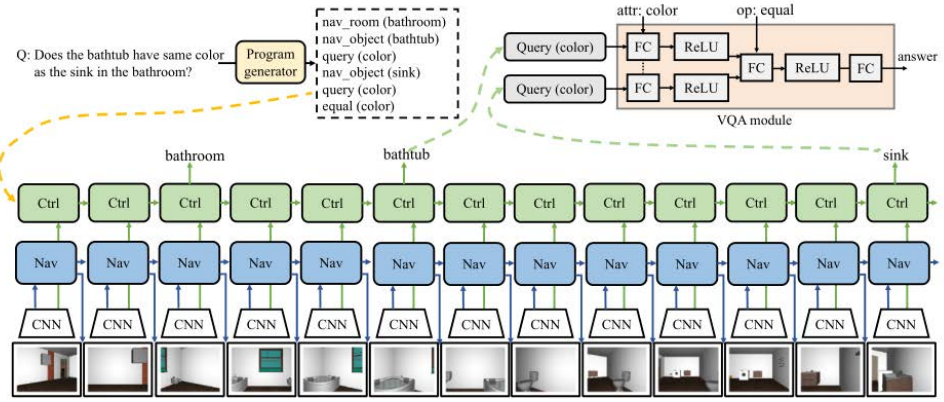
- Neural Modular Control for Embodied Question Answering. CoRL, 2018.

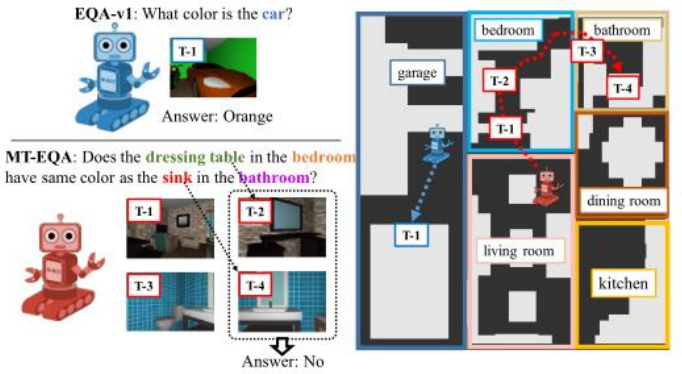
4 具身语言问答

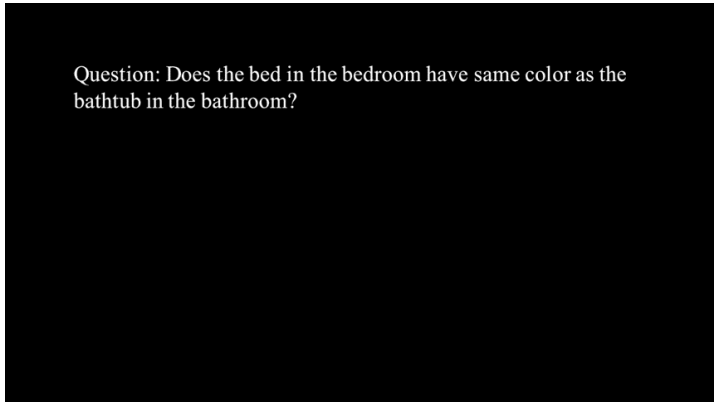
➢ 前沿：多目标问答

Multi-Target Embodied Question Answering

MT-EQA的问题涉及到需要导航的多个目标(如卧室、梳妆台、浴室、水槽)，以及多个目标(如梳妆台、水槽)之间的属性比较。将原来的EQA问题从有限的单目标设置扩展到更具挑战性的多目标设置，要求智能体在回答问题之前进行比较推理





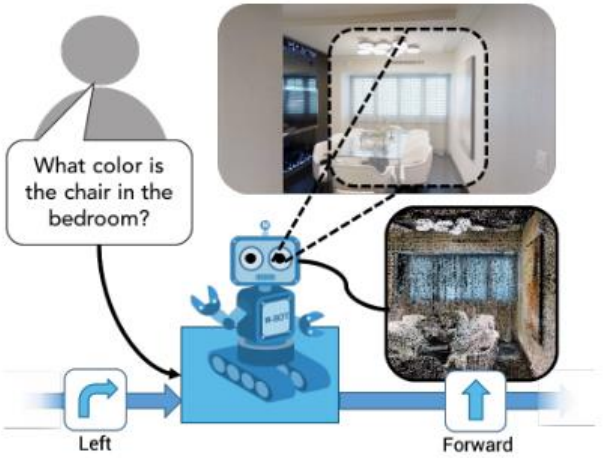


- Multi-target embodied question answering. CVPR, 2019.

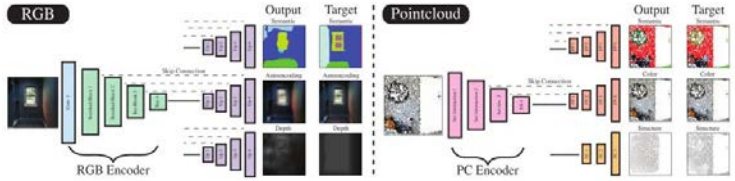
4 具身语言问答

➢ 前沿：引入点云

Embodied Question Answering in Photorealistic Environments with Point Cloud Perception



仿真环境：Matterport 3D
数据集：MP3D-EQA dataset
发现：点云为学习避障提供了比RGB图像更丰富的信息
网络：开发了一个端到端的可训练导航模型，能够直接从3D点云学习目标驾驶导航策略。设计了两种网络，一个简单的前馈网络，它将最近的5个视觉观察结果的嵌入串联起来作为输入来预测一个动作。第二个是一个两层的GRU+RNN，将当前观察和之前动作的编码作为输入来预测当前动作。通过网络与输入组合进行对比（网络的输入为RGB, Pointcloud，RGB+Pointcloud），证明点云信息对导航更有效



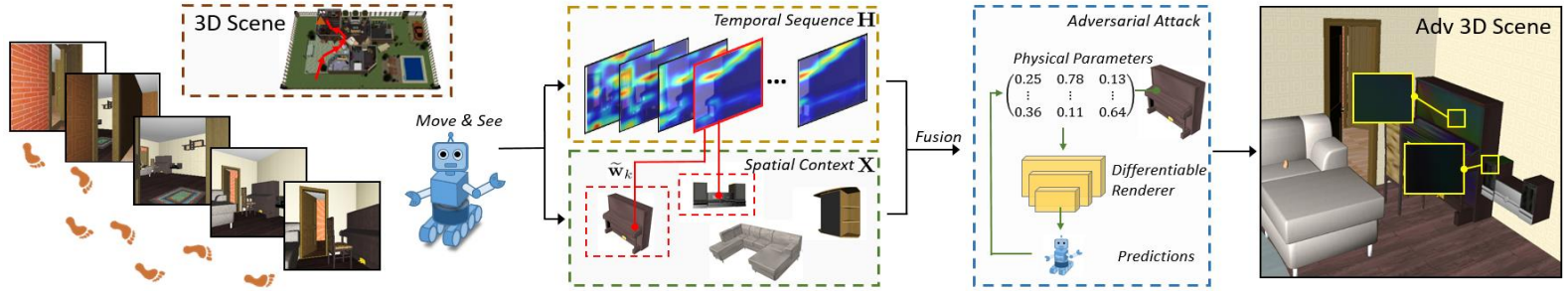
- Embodied question answering in photorealistic environments with point cloud perception. CVPR, 2019.

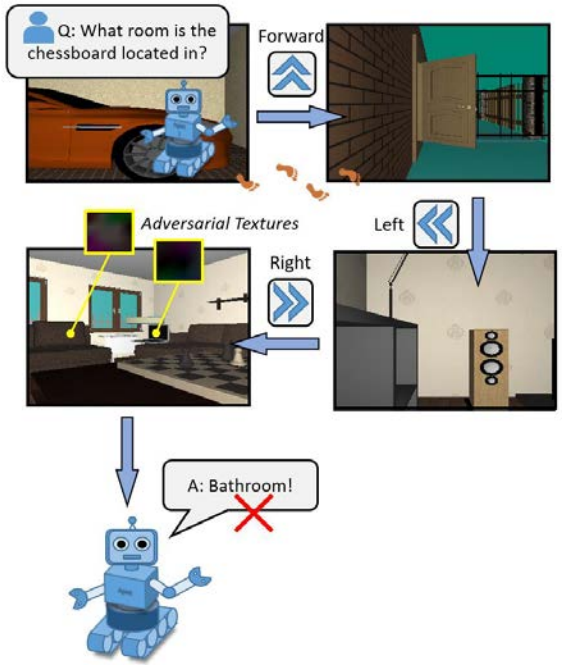
4 具身语言问答

➢ 前沿：对抗

新浪网：“忽悠”智能机器人，竟然改改物品纹理就成功了
https://tech.sina.com.cn/roll/2020-07-14/doc-iivhuipn2860336.shtml

- 问题背景
  - I. 大量的对抗攻击研究都关注于静态场景，而对于具身机器人的鲁棒性仍未有相关研究；
  - II. 对抗攻击有助于我们提升对具身机器人的理解并进一步提升其鲁棒性。





- 方法与效果
  - I. 提出了时空融合的对抗攻击算法，生成人眼不可区分的对抗纹理，攻击具身机器人；
  - II. 在EQA和EVR等任务上成功攻击具身机器人。

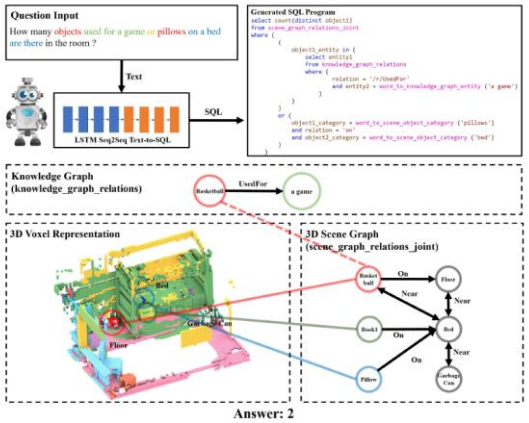
- 研究意义
  - I. 提升对深度学习模型的理解（texture vs shape）；
  - II. 通过对抗训练可以增强模型在高斯噪音等环境中的泛化能力。

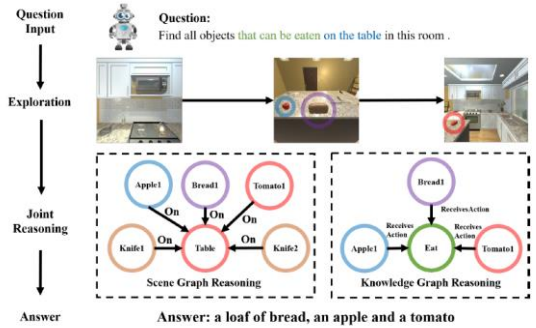
	QA		Navigation	
	Adv	Gaussian	Adv	Gaussian
Vanilla	5.67%	22.14%	1.39	1.20
AT	23.56%	38.87%	1.17	1.01
GT	8.49%	32.90%	1.32	1.09

Liu, Liu, and Tao, “Spatiotemporal Attacks for Embodied Agents”, ECCV, 2020.

4 具身语言问答

➢ 前沿：引入知识





- Knowledge-based Embodied Question Answering, T-PAMI, 2023

4 具身语言问答

➢ 前沿：交互问答

- 交互式问题回答(IQA)，即回答问题的任务，需要一个智能体与动态视觉环境交互。IQA向智能体提出了一个场景和一个问题，比如:“冰箱里有苹果吗?”智能体必须在场景中导航，获得对场景元素的视觉理解，与对象互动(例如打开冰箱)，并根据问题计划一系列行动。
- 仿真环境： AI2thor
- 数据集：IQUAD V1

IQA: Visual Question Answering in Interactive Environments

Daniel Gordon[2] Aniruddha Kembhavi[1] Mohammad Rastegari[1]
Joseph Redmon[2] Dieter Fox[2] Ali Farhadi[1,2]
[1]Allen Institute for Artificial Intelligence
[2]Paul G. Allen School of Computer Science, University of Washington

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

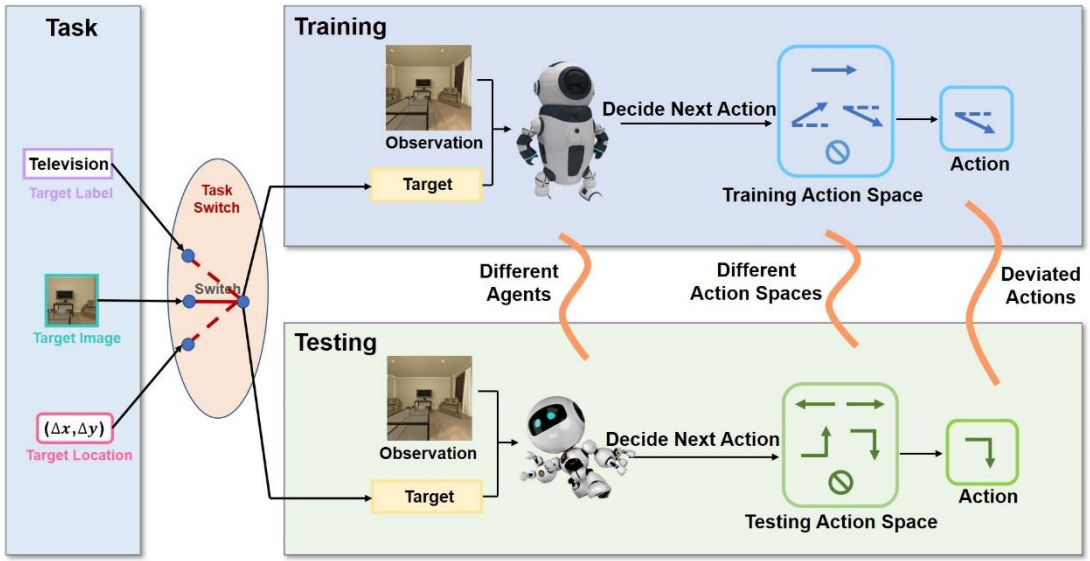
- 与Embodied QA区别：
  - Embodied QA要求智能体导航到目标附近，可以“看到”目标，通过“看到的”像素进行回答
  - IQA要求智能体导航到目标附近，并且对场景中的对象进行交互，最终“看到”目标进行回答

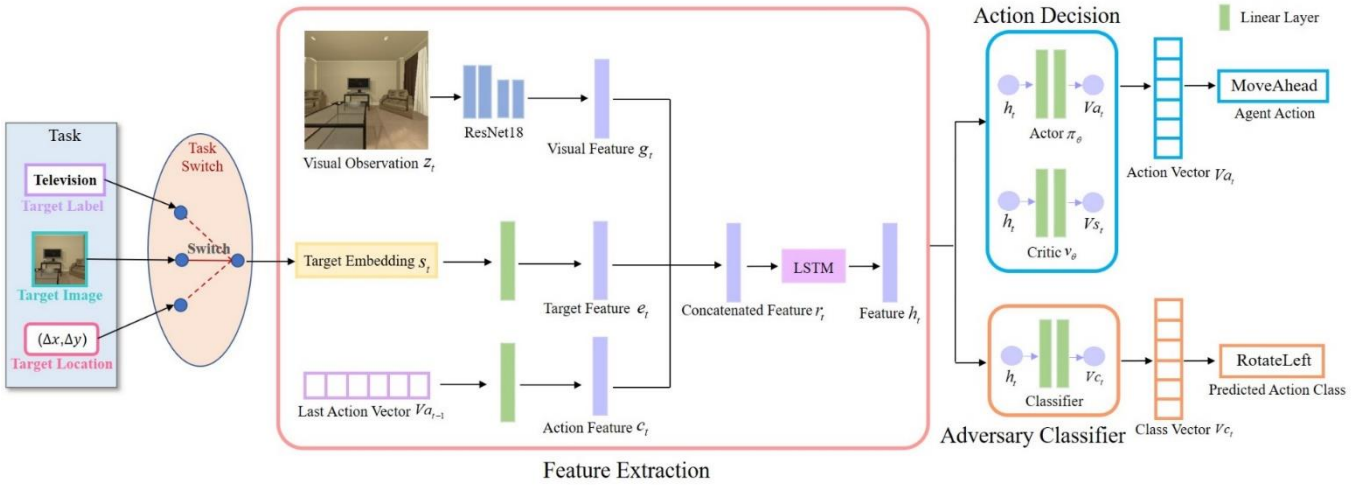
- IQA: Visual question answering in interactive environments. CVPR, 2018

4 具身语言问答

➢ 前沿：具身失配

- 对抗训练 —— 学习鲁棒的状态特征表示
  - 特征提取：提取视觉、目标以及历史动作特征，融合后得到状态特征
  - 动作决策：Actor-Critic结构生成下一步动作
  - 对抗学习：预测产生当前状态的前一步动作，希望对抗学习层无法分辨前一步动作类型
- 适应训练 —— 策略迁移至新的动作空间
  - 更新Actor网络维数，使用较少的样本训练适应于新测试动作空间的策略
- 该模块可插入不同的导航决策模型中





4 具身语言问答

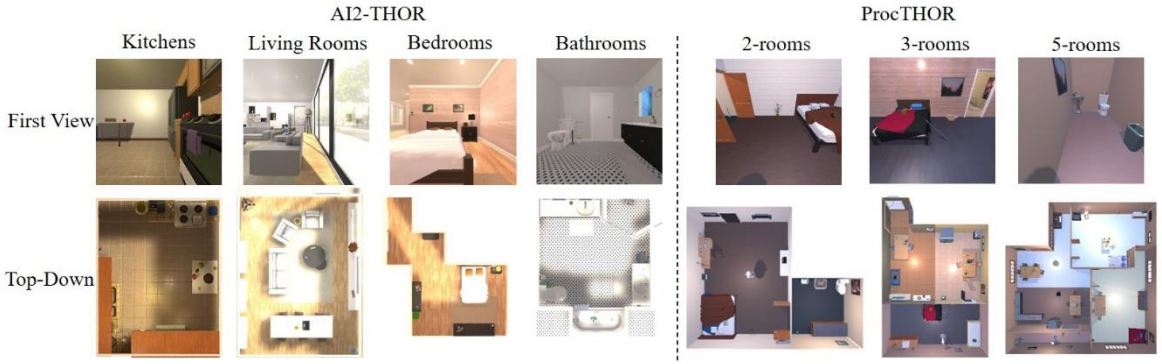
➢ 前沿：具身失配

- 实验数据集
  - **AI2-THOR**
  - **Sub-Proc (ProcTHOR-10K子集)**
- 测试动作空间（5种设置）

  $A_{test} = A_{train}$, $A_{test} \subseteq A_{train}$ $A_{test} \supseteq A_{train}$ $A_{test} \cap A_{train} \neq \varnothing$ $A_{test} \cap A_{train} = \varnothing$
- 评价指标
  - 成功率 $SR = \frac{1}{N}\sum_{i=1}^{N} R_i$
  - 路径加权成功率 $SPL = \frac{1}{N}\sum_{i=1}^{N} R_i \frac{L_i}{\max(L_i, G_i)}$

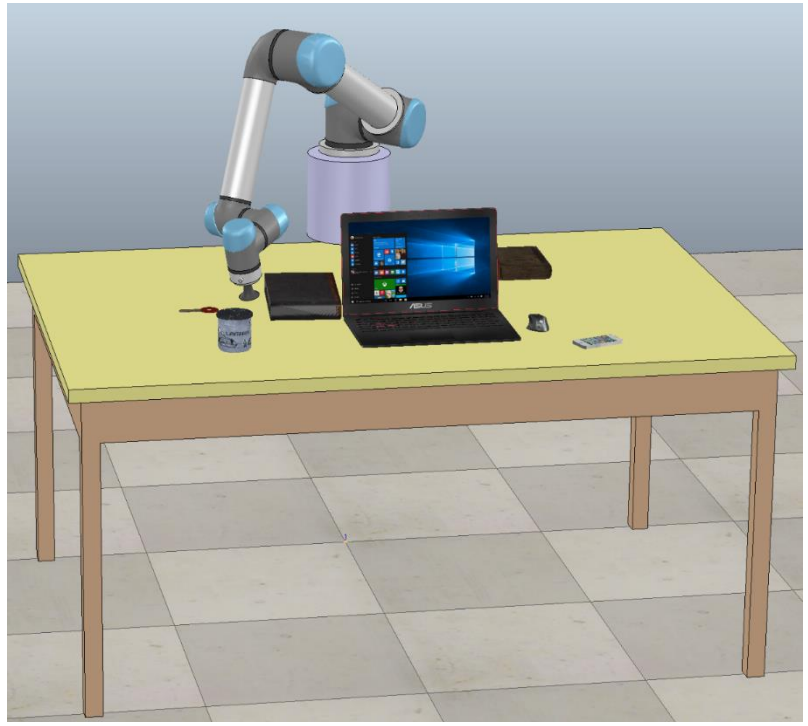
QUANTITATIVE RESULTS OF THE PROPOSED MODEL MODEL IN TASK *VSN-Label* IN AI2-THOR AND SUB-PROC

Dataset	Methods	$A_{test} = A_{train}$ SR(%)	$A_{test} = A_{train}$ SPL(%)	$A_{test} \subseteq A_{train}$ SR(%)	$A_{test} \subseteq A_{train}$ SPL(%)	$A_{train} \subseteq A_{test}$ SR(%)	$A_{train} \subseteq A_{test}$ SPL(%)	$A_{train} \cap A_{test} \neq \varnothing$ SR(%)	$A_{train} \cap A_{test} \neq \varnothing$ SPL(%)	$A_{train} \cap A_{test} = \varnothing$ SR(%)	$A_{train} \cap A_{test} = \varnothing$ SPL(%)
AI2-THOR	Random	6.70	3.21	6.60	3.25	6.30	3.16	7.00	3.41	6.50	2.91
	A3C	39.50	12.56	29.60	9.90	29.80	9.16	28.30	9.55	28.00	9.67
	ScenePriors	44.70	17.25	34.80	12.73	33.90	14.03	33.70	11.10	33.40	11.70
	SAVN	47.50	13.62	39.20	10.63	38.80	11.50	37.60	10.56	37.60	12.80
	SemMap	53.20	18.65	45.60	14.23	44.10	13.69	43.30	13.37	43.50	13.66
	Noise-A3C	40.90	13.95	31.70	11.69	32.40	11.23	30.90	10.88	30.60	10.75
	Noise-ScenePriors	47.30	19.42	37.40	14.45	37.10	14.87	36.70	13.33	36.40	13.32
	Noise-SAVN	49.50	15.67	41.30	13.90	41.00	13.25	40.80	13.01	40.50	12.85
	Noise-SemMap	56.50	20.12	48.10	15.78	47.80	15.26	47.60	15.19	47.20	14.63
	EMAL-A3C	49.00	15.56	46.20	13.51	46.40	11.55	44.90	13.06	44.90	11.31
	EMAL-ScenePriors	57.70	**31.83**	55.90	15.91	56.20	15.03	55.30	**18.69**	54.90	17.55
	EMAL-SAVN	60.30	16.33	58.20	15.72	58.40	13.93	58.10	13.61	57.60	12.88
	EMAL-SemMap	**66.70**	25.39	**64.30**	**20.08**	**64.10**	**18.29**	**63.90**	18.05	**63.60**	**17.63**
Sub-Proc	Random	1.20	0.65	1.20	0.60	0.90	0.53	1.40	0.83	1.00	0.62
	A3C	33.60	9.23	23.10	6.65	23.60	6.39	22.50	6.58	22.20	6.16
	ScenePriors	38.20	14.17	28.80	9.65	28.60	9.88	27.90	8.25	27.60	8.13
	SAVN	42.30	12.67	34.90	9.09	33.20	8.99	33.00	8.63	32.70	8.55
	SemMap	48.10	15.98	40.60	11.16	40.10	10.98	39.80	10.76	39.90	10.29
	Noise-A3C	34.50	9.87	24.90	8.19	25.10	8.33	24.10	7.98	24.00	7.83
	Noise-ScenePriors	41.20	16.05	31.20	11.35	31.40	11.23	30.80	11.09	30.60	11.01
	Noise-SAVN	43.60	12.39	35.90	10.88	36.00	10.56	35.10	10.23	35.00	10.09
	Noise-SemMap	50.30	17.01	43.30	12.17	43.20	12.09	42.80	11.86	42.50	11.59
	EMAL-A3C	43.30	13.09	40.10	10.26	40.60	10.01	39.70	9.82	39.20	9.67
	EMAL-ScenePriors	51.30	**27.75**	49.70	13.21	49.50	13.23	49.20	12.93	49.00	12.67
	EMAL-SAVN	54.10	14.33	52.30	14.04	52.40	13.15	52.10	12.87	52.00	12.65
	EMAL-SemMap	**61.30**	22.12	**59.60**	**19.86**	**59.70**	**18.05**	**59.30**	**17.76**	**59.10**	**17.33**



4 具身语言问答

➢ 前沿：操作问答



How about robots?
Can robots deal with this kind of QA like humans?

- MQA: Answering the question via robotic manipulation, *RSS*, 2021

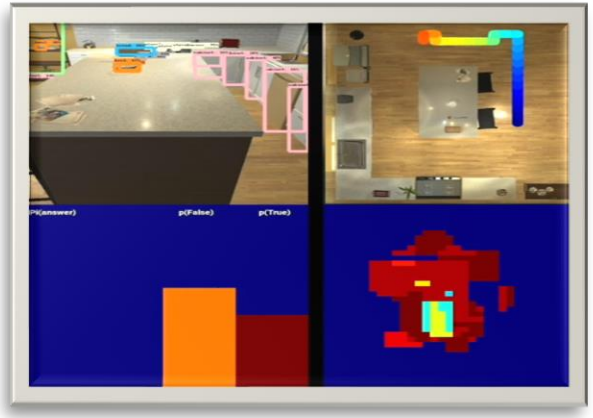
4 具身语言问答

➢ 前沿：操作问答

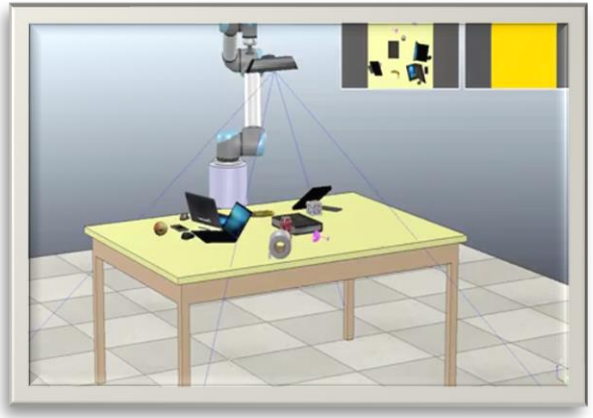
Visual Question Answering (VQA)

Embodied Question Answering (EQA)

	VQA	EQA	IQA	MQA
Understanding	✓	✓	✓	✓
Exploration	–	✓	✓	✓
Interaction	–	–	✓	✓
Manipulation	–	–	–	✓



Interactive Question Answering (IQA)



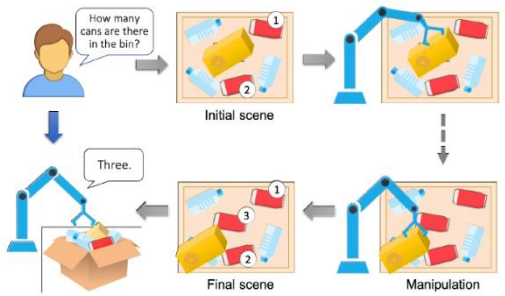
Manipulation Question Answering (MQA)

A comprehensive comparison of VQA, EQA, IQA and the proposed MQA tasks is illustrated in the above table.

- MQA: Answering the question via robotic manipulation, *RSS*, 2021

4 具身语言问答

➢ 前沿：操作问答



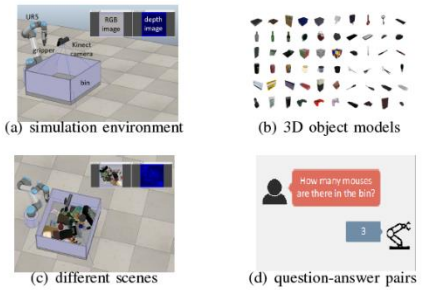
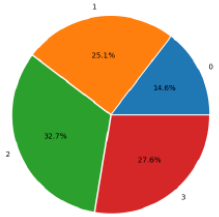
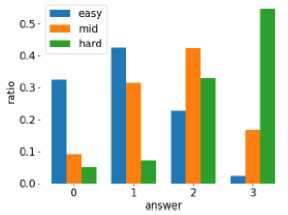
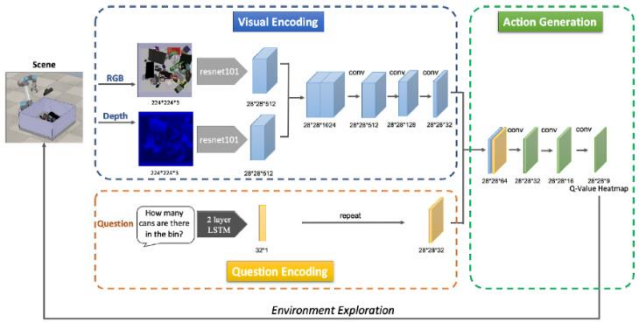
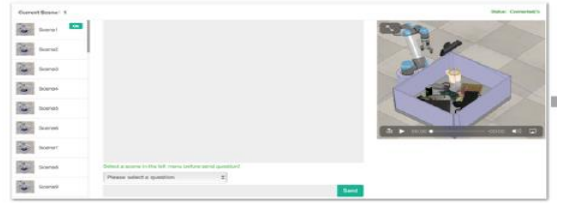


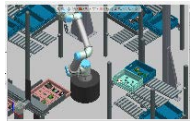
Fig. 2. The overview of the MQA dataset.









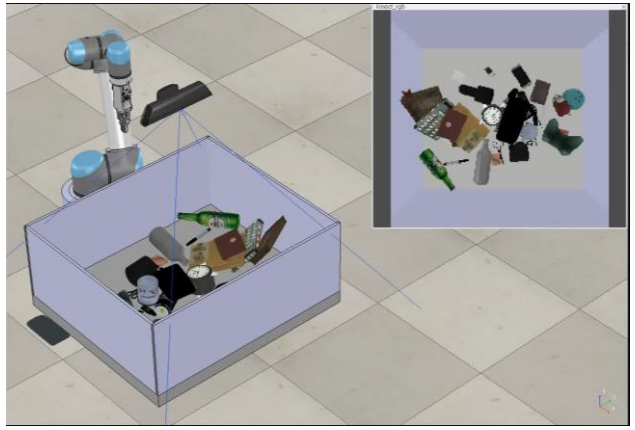


- MQA: Answering the question via robotic manipulation, *RSS*, 2021

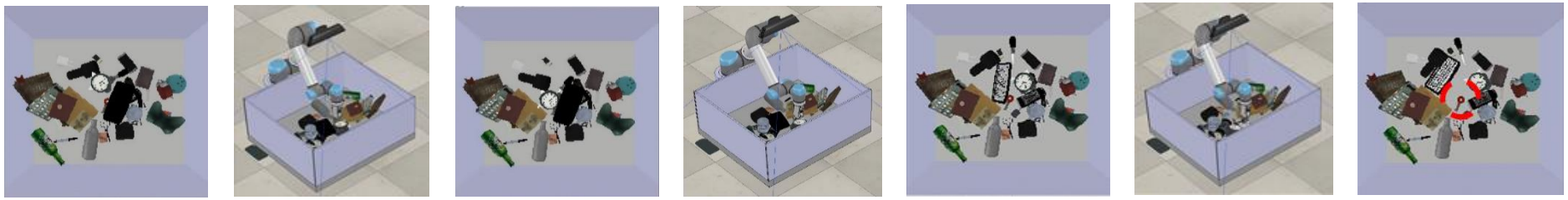
4 具身语言问答

➢ 前沿：操作问答

question: Is there a key in the bin?



state 0 manipulation state 1 manipulation State 2 manipulation state 3

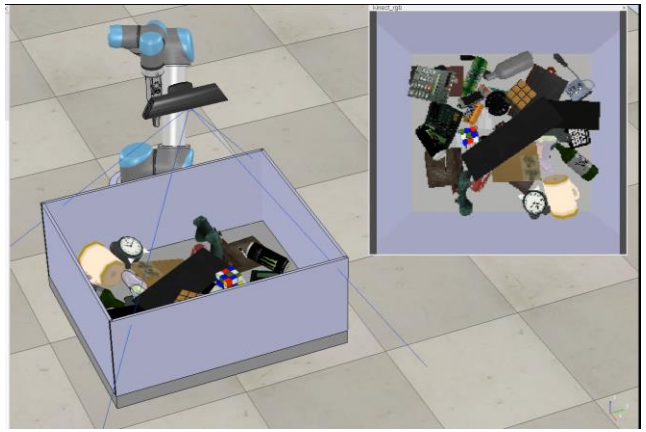


- MQA: Answering the question via robotic manipulation, *RSS*, 2021

4 具身语言问答

➢ 前沿：操作问答

question: How many keyboards are there in the bin?

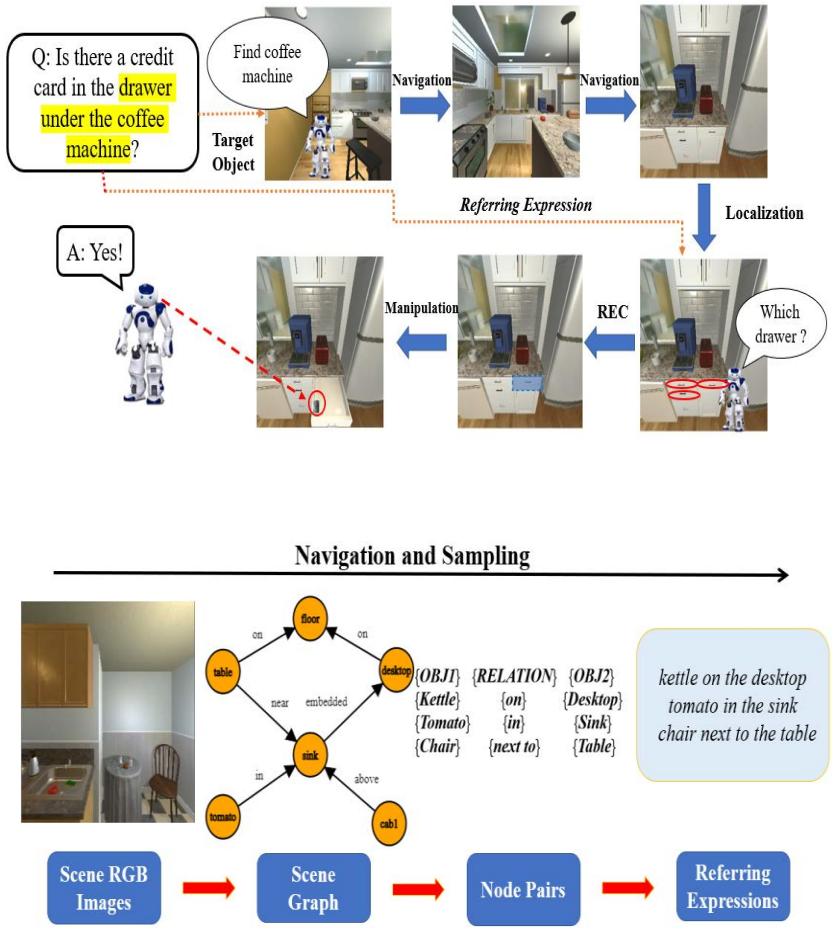


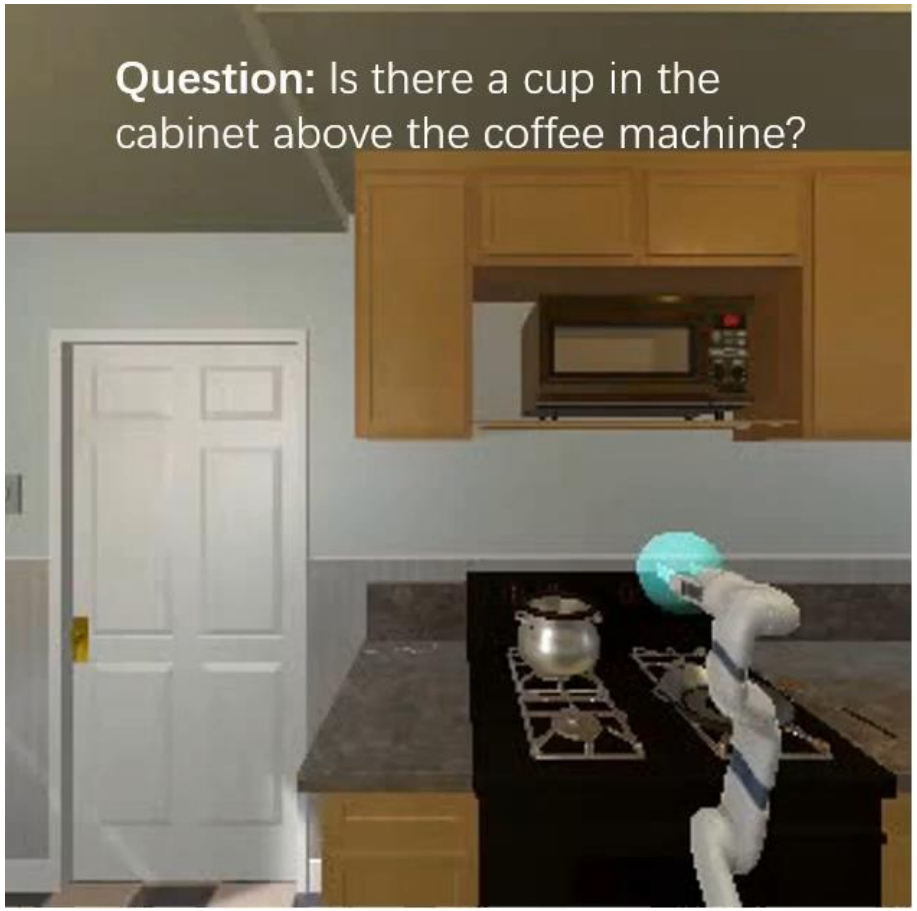
state 0 manipulation state 1 manipulation state 2

- MQA: Answering the question via robotic manipulation, *RSS*, 2021

4 具身语言问答

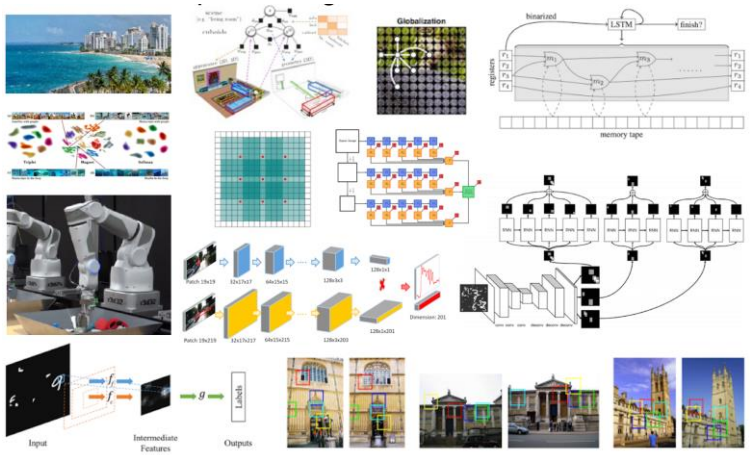
➢ 前沿：操作问答



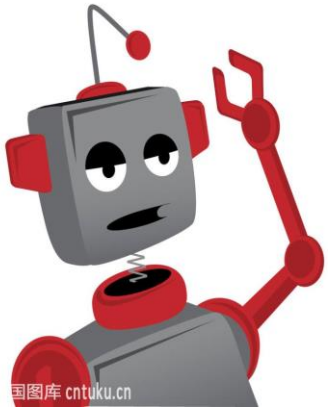


- Embodied Referring Expression for Manipulation Question Answering in Interactive Environment, ICRA, 2023

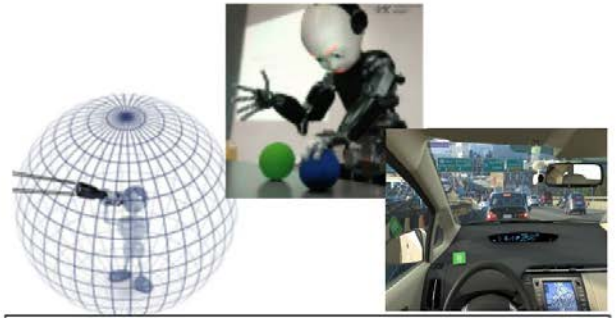
总结与展望



主动感知



disembodied bags of manually labeled images



embodied learning by self-supervision

总结与展望

具身智能

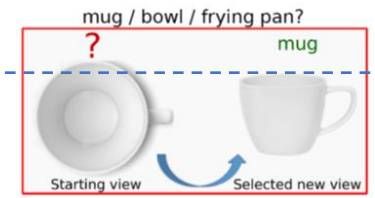
模式识别
机器学习
基础模型
多模态
（视觉，语言，语音）

互联网

物理世界

视觉测量
视觉伺服
视觉导航
多模态
（视觉，触觉，听觉）

主动感知



感知-理解

感知-行为

M.J.Tarr M.J.Black

R.Bajcsy J.Y.Aloimonos

1994

计算机视觉

智能机器人

工业机器人