业界具身
技术路线解析

时事热点

智能体 | RAG | 具身智能 | 智能驾驶 | 工业大模型 | …

大模型训推

6. 大模型数据&算法

- **数据&模型评估**：Prompt 工程，模型评估算法和测评体系
- **大模型算法**：Scaling Law、Transform 结构，LLM/MLM 模型

7. 大模型训练

- **分布式训练**：TP/DP/PP/SP/EP 并行，Megatron、DeepSpeed 分布式并行库介绍
- **微调**：全参微调、底参微调(LoRA/QLoRA 等)、指令微调

8. 大模型推理

- **推理框架**：VLLM、推理框架的架构，推理框架线程池等构架
- **推理优化**：大模型推理加速(XXXAttention)、长序列推理优化算法

编译计算架构

4. 计算架构

- **传统编译器**：传统编译器 GCC与LLVM
- **AI 编译器**：AI编译器发展与架构定义，未来挑战与思考
- **前端优化**：前端优化（算子融合、内存优化等）
- **后端优化**：后端优化(Kernel优化、Auto Tuning)
- **多面体**：复杂的循环依赖关系映射到高维几何空间

5. 通信架构

- **集合通信**：通信原语、通信原理、集合通信算法
- **NCCL/HCCL**：集合通信库、网络拓扑、通信方式、通信算法，NCCL 架构

硬件体系结构

3. AI 集群

- **集群管理运维**：K8s集群运维、K8s容器、集群监控等工具
- **集群性能指标**：稳定性、吞吐、线性度等
- **集群训推一体化**：训练、推理大模型执行，训练推理显存分析
- **机房建设**：风火水电、夜冷、柜板等知识

1. AI 芯片

- **AI 计算体系**：AI 计算模式与计算体系架构
- **AI 芯片基础**：CPU、GPU、NPU等芯片体基础原理
- **英伟达GPU**：英伟达GPU TensorCore、NVLink剖析
- **国外AI芯片**：谷歌、特斯拉等专用AI处理器核心原理
- **国内AI芯片**：寒武纪、燧原科技等专用AI处理器原理

2. 通信与存储

- **通信**：路由器、交换机基本原理和网络拓扑
- **存储**：DRAM、SRAM、存储 POD 到大模型存储 CKPT 算法

具身智体的典型架构

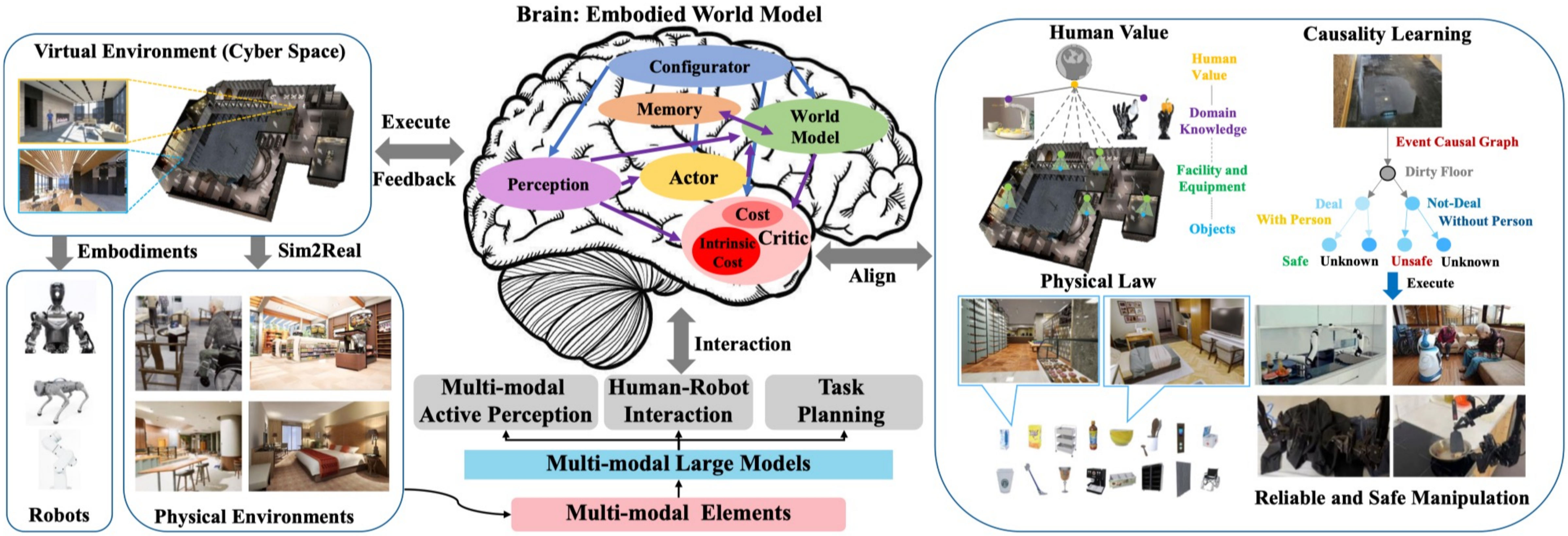


Fig. 2. The overall framework of the embodied agent based on MLMs and WMs. The embodied agent has a embodied world model as its “brain”. It has the capability to understand the virtual-physical environment and actively perceive multi-modal elements. It can fully understand human intention, align with human value and event causality, decompose complex tasks, and execute reliable actions, as well as interact with humans and utilize knowledge and tools.

具身智能涉及知识点

- 具身机器人；具身模拟器；具身感知；具身交互；具身智体；模拟到现实，包括具身WM、数据以及控制。

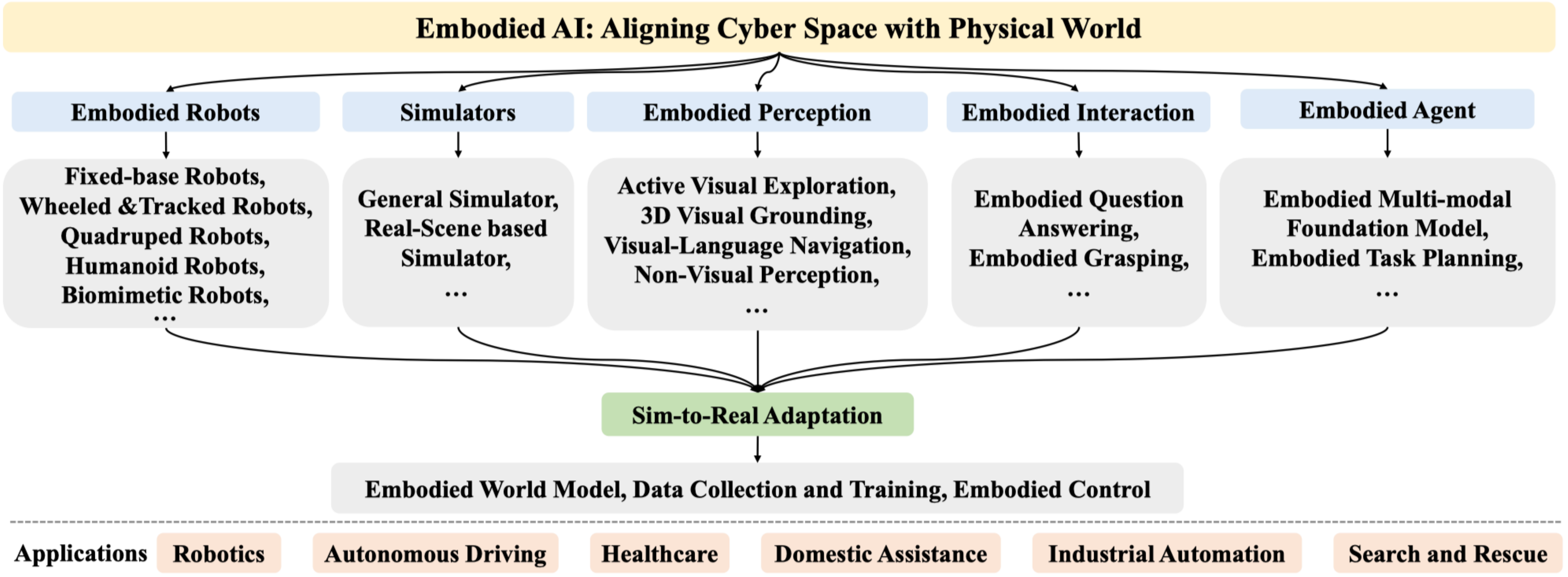


Fig. 3. This survey focuses on comprehensive analysis of the latest advancements in embodied AI.

具身智能 Embodied Intelligence

- 具身智能（Embodied Intelligence）高级机器智能形式，它使机器人能够像人类一样感知和理解环境，并通过自主学习和适应性行为来完成任务。机器人的能力和实现过程抽象为：

感知 - 决策 - 执行

01 具身感知

具身感知：理解场景、预测和执行

- 传统机器人的模式识别主要识别图像中的目标。具身感知的智体必须在物理世界中移动并与环境互动，需要对 3D 空间和动态环境有更透彻的理解。
- 具身感知需要具备视觉感知和推理能力，理解场景中的三维关系，并基于视觉信息预测和执行复杂任务。

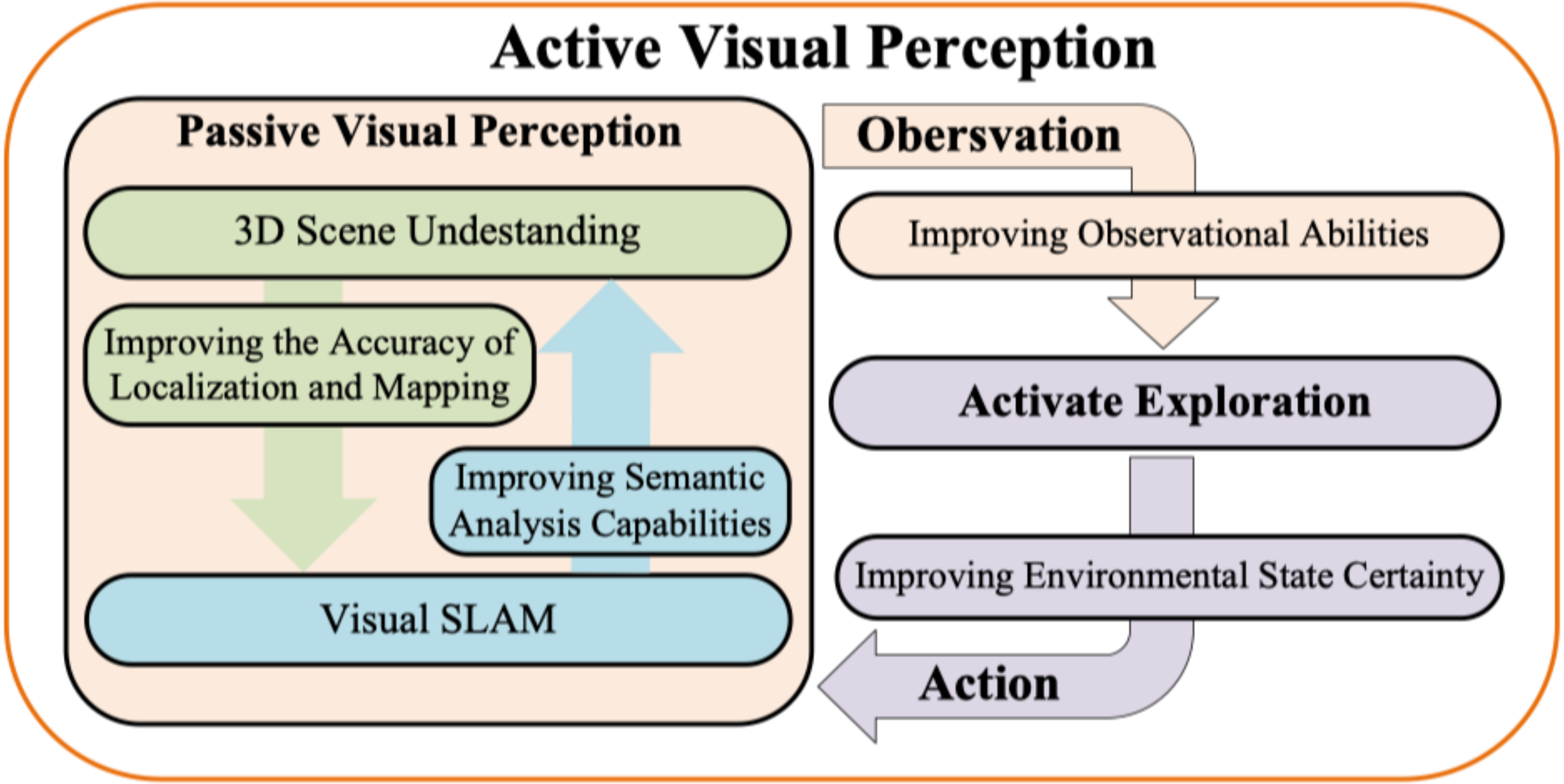
具身感知：理解场景、预测和执行

- 传统机器人的模式识别主要识别图像中的目标。具身感知的智体必须在物理世界中移动并与环境互动，需要对 3D 空间和动态环境有更透彻的理解。
- 具身感知需要具备视觉感知和推理能力，理解场景中的三维关系，并基于视觉信息预测和执行复杂任务。

主动感知

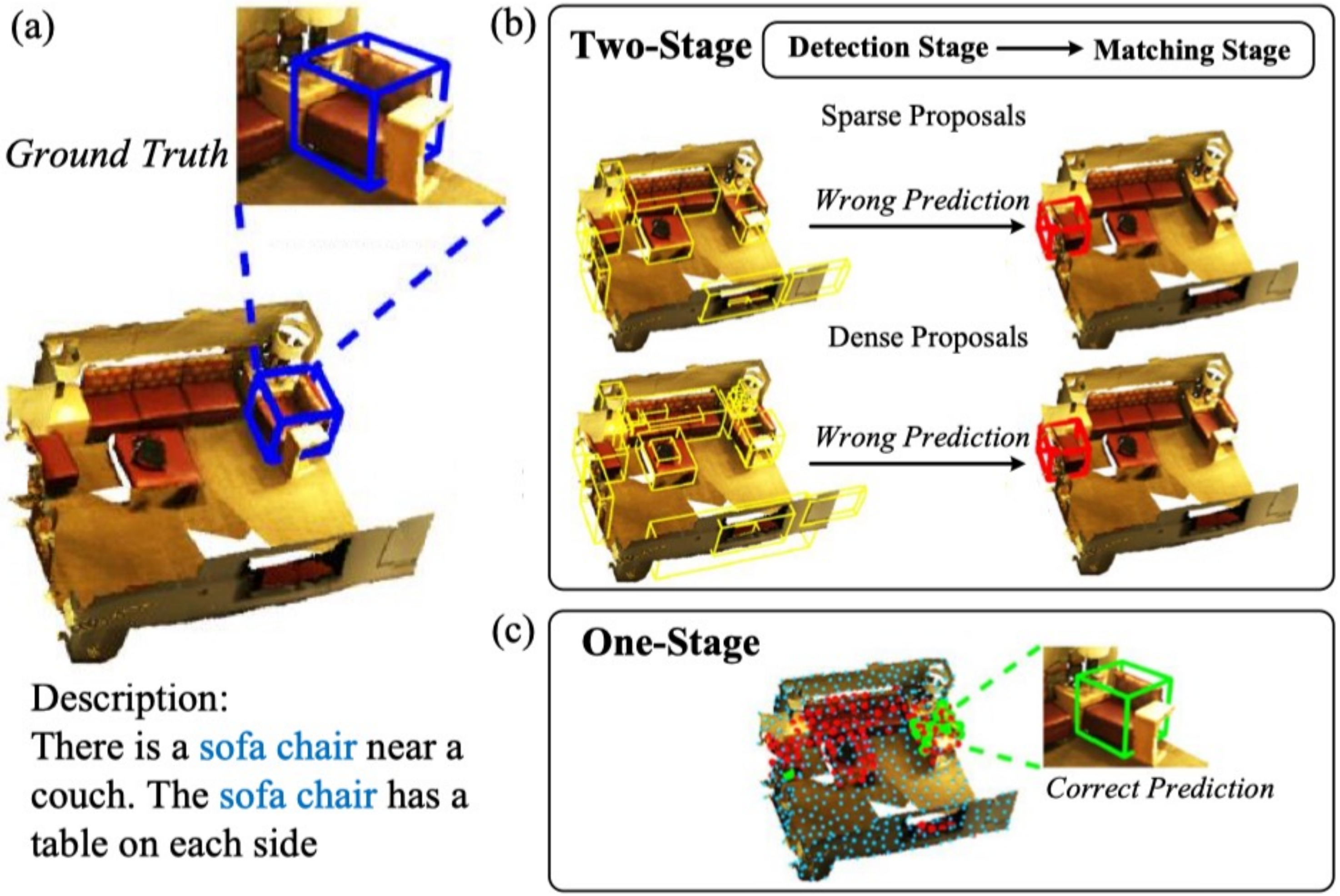
被动感知

具身主动感知：视觉SLAM、3D场景理解、主动探索



具身被动感知：3D视觉 Grounding

1. 二阶段3D Grounding
1. 一阶段 3D Grounding



具身被动感知：视觉语言导航 Visual Language Navigation

- VLN 要求机器人理解复杂多样的视觉观察，同时解释不同粒度的指令。视觉信息可以是过去轨迹的视频，也可以是一组历史当前观测图像。

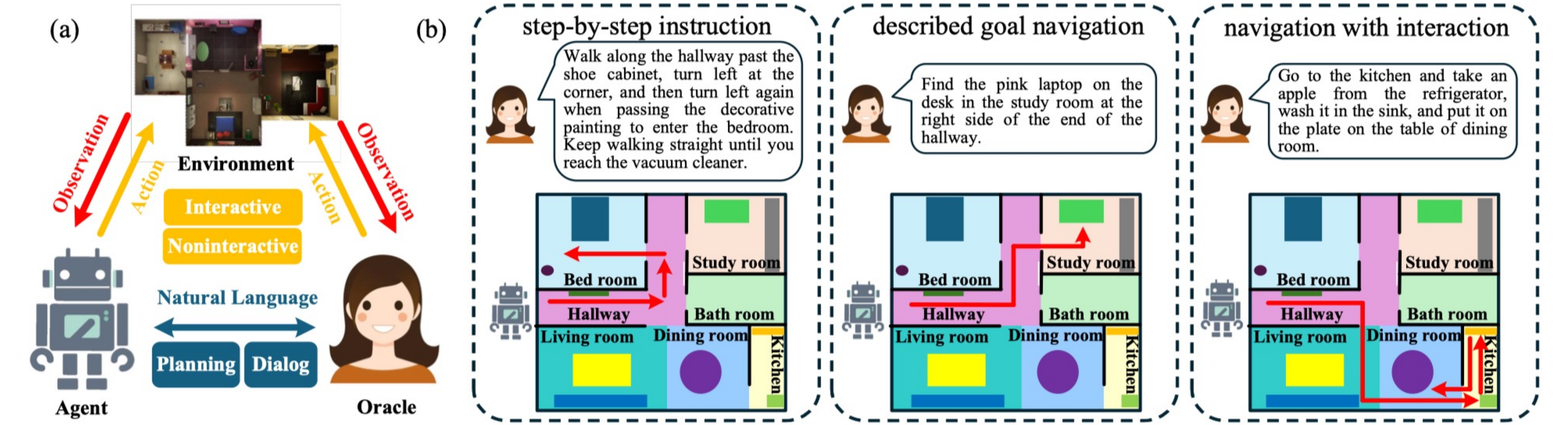


Fig. 9. (a) Overview of VLN. The embodied agent communicates with humans through natural language. Humans issue instructions to the embodied agent, who completes tasks such as planning and dialog. Subsequently, through collaborative cooperation or the embodied agent's independent actions, actions are made in interactive or non-interactive environments based on visual observations and instructions, (b) Different tasks of VLN.

02. 具身决策

aka 具身智体

具身智体的典型架构

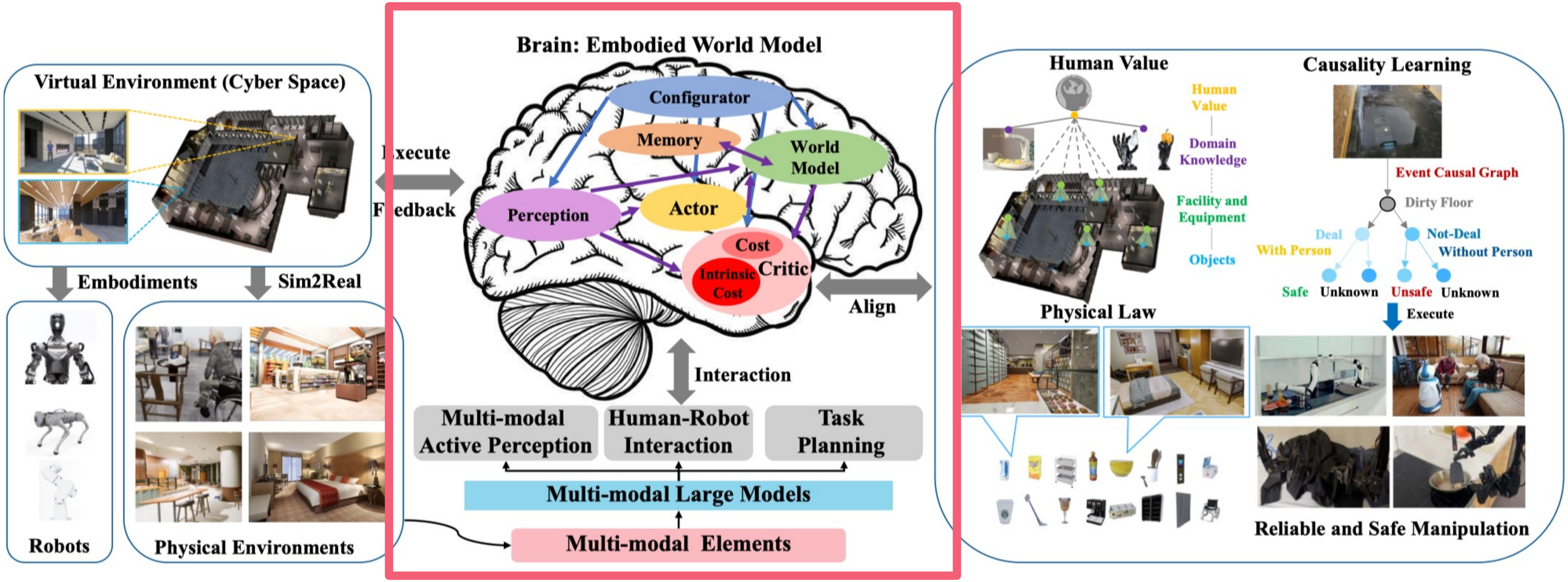
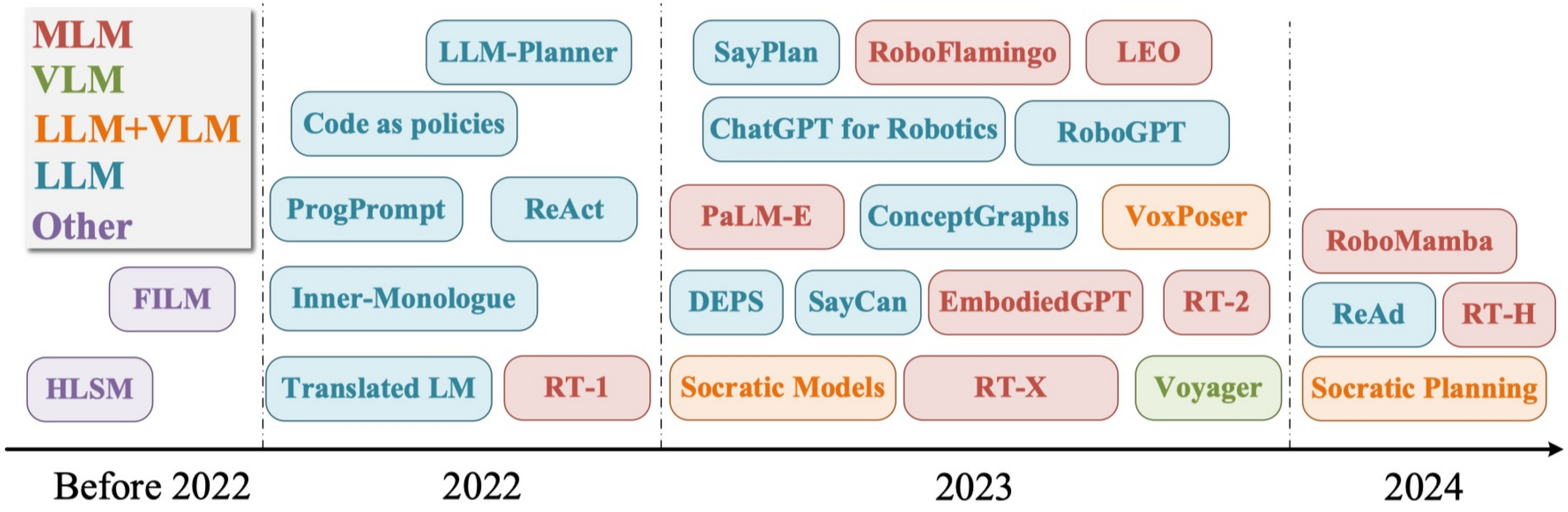


Fig. 2. The overall framework of the embodied agent based on MLMs and WMs. The embodied agent has a embodied world model as its "brain". It has the capability to understand the virtual-physical environment and actively perceive multi-modal elements. It can fully understand human intention, align with human value and event causality, decompose complex tasks, and execute reliable actions, as well as interact with humans and utilize knowledge and tools.

具身决策

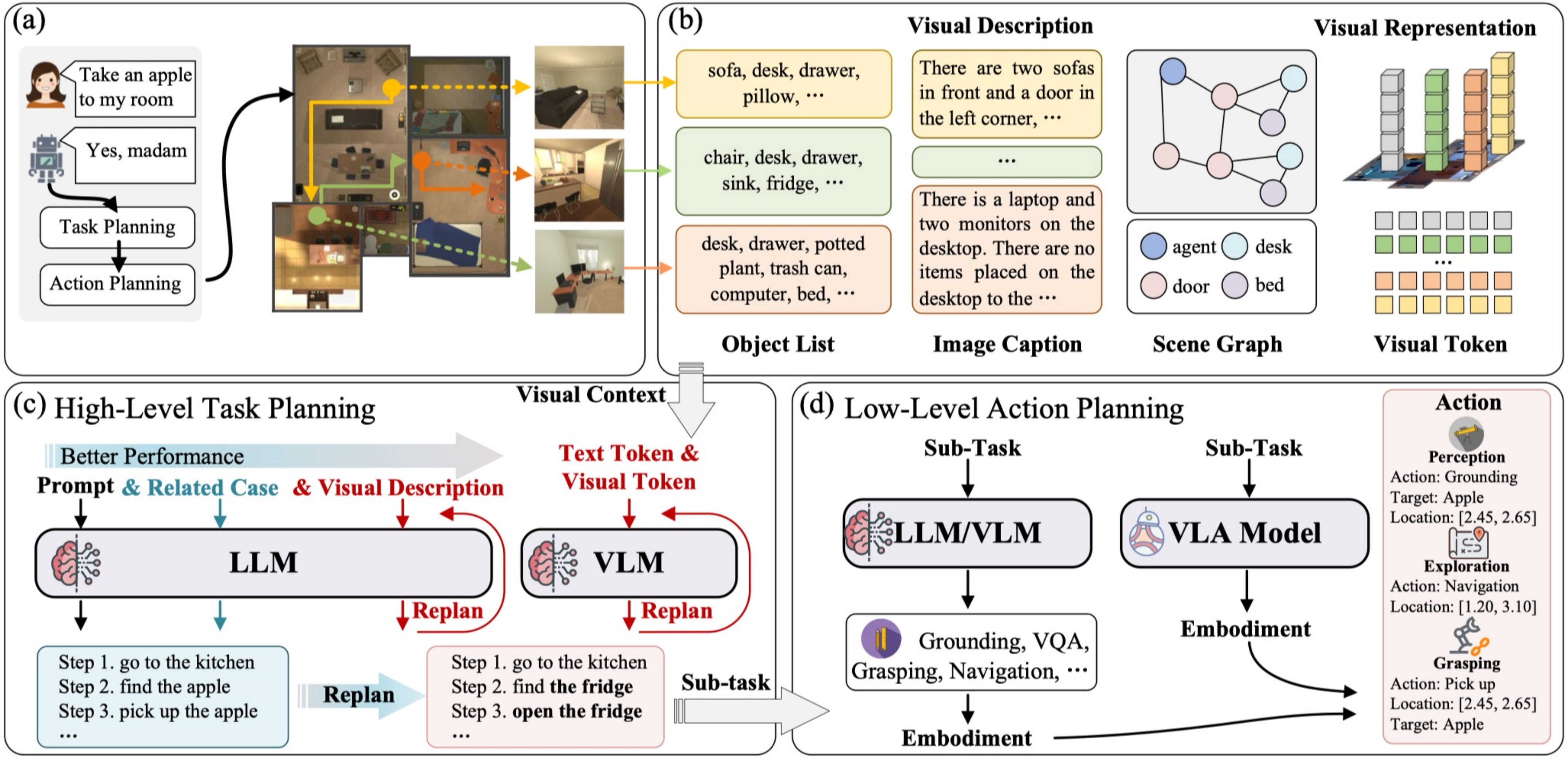
- 智体被定义为能够感知环境并采取行动以实现特定目标的自主实体。大模型进一步扩大了智体在实际场景中的应用。
- 多模态大模型的智体被具身化为物理实体时，能够有效地将大模型能力从虚拟空间转移到物理世界，从而成为具身智体。



具身决策

- 为了完成任务，具身智体通常涉及以下过程：
  1. 将抽象而复杂的任务分解为具体子任务，即高层具身任务规划。
  2. 利用具身感知和具身交互模型，逐步实施子任务，即低层具身行动规划。

具身决策



(a)
Take an apple to my room
Yes, madam
Task Planning
Action Planning
(b)
Visual Description
Visual Representation
sofa, desk, drawer, pillow, ···
chair, desk, drawer, sink, fridge, ···
desk, drawer, potted plant, trash can, computer, bed, ···
There are two sofas in front and a door in the left corner, ···
···
There is a laptop and two monitors on the desktop. There are no items placed on the desktop to the ···
agent
desk
door
bed
Object List
Image Caption
Scene Graph
Visual Token
Visual Context
(c) High-Level Task Planning
Better Performance
Prompt & Related Case & Visual Description
Text Token & Visual Token
LLM
VLM
Replan
Replan
Step 1. go to the kitchen
Step 2. find the apple
Step 3. pick up the apple
···
Replan
Step 1. go to the kitchen
Step 2. find the fridge
Step 3. open the fridge
···
Sub-task
(d) Low-Level Action Planning
Sub-Task
Sub-Task
LLM/VLM
VLA Model
Grounding, VQA, Grasping, Navigation, ···
Embodiment
Embodiment
Action
Perception
Action: Grounding
Target: Apple
Location: [2.45, 2.65]
Exploration
Action: Navigation
Location: [1.20, 3.10]
Grasping
Action: Pick up
Location: [2.45, 2.65]
Target: Apple

03. 具身执行

aka 具身交互

具身执行 or 具身交互

- 智体在物理或模拟空间中与人类和环境互动的场景，采取具体执行的动作。

任务问答

具身抓取

具身交互：任务问答

- 智体在物理或模拟空间中与人类和环境互动的场景。e.g. 具身问答任务中，智体需要从第一人称视角探索环境，收集回答问题所需的信息；
- 具有自主探索和决策能力的智体，不仅要考虑采取哪些行动来探索环境，还需决定何时停止探索以回答问题。

具身交互：任务问答

Exploring...
turn right
turn right
turn right
move forward
turn left
move forward
move forward
Knowledge-based
Question: Is there an object in the living room used to lower the temperature?
Answer: Yes.
Episodic Memory
Episodic History
Question: Where is the clock?
Answer: It is mounted on the wall above the cabinets, directly across from the stove.
Single Objective
Question: What room is the vase located in?
Answer: Living room.
Multiple Objectives
Question: Is the dinner table the same color as the TV cabinet?
Answer: No.
Multi-agent
Question: Is there a washing machine in the house?
: A1
: A2
: A3
ANSWER
Answer: Yes.
Interaction
Question: Are there any apples in the fridge?
Action: open the fridge
Answer: No.
Object States
Question: Could someone be watching TV in the living room?
TV on
air conditioner on
curtain withdrawn
Answer: Yes.

具身交互：具身抓取

- 根据人类指令执行操作，如抓取、放置目标；需要语义理解、场景感知、决策和鲁棒控制规划。
- 具身抓取方法将传统机器人运动学抓取与 LLM/VLM/MLM 等大模型结合，使智体能够在多感知器下执行抓取任务。

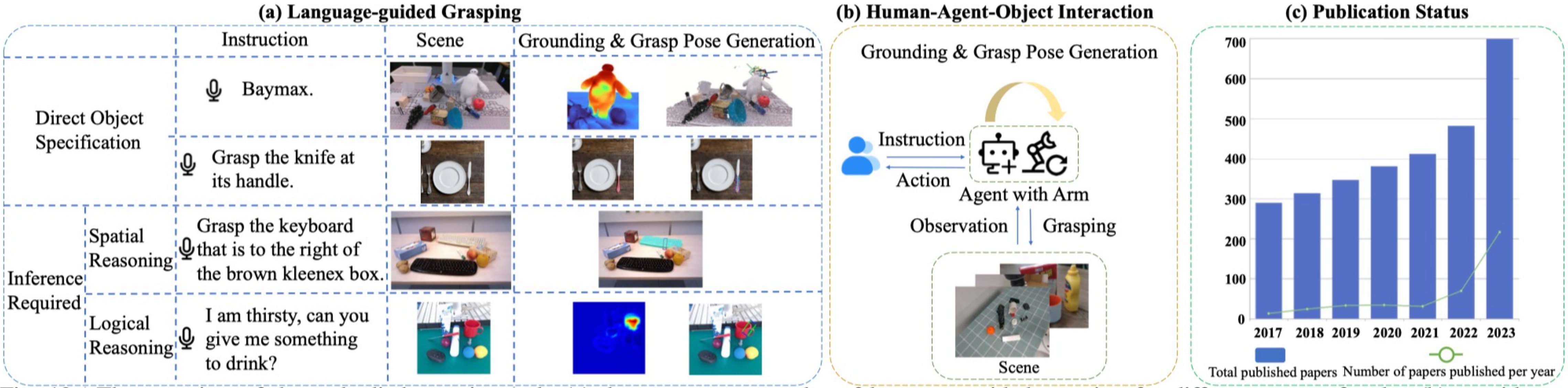


Fig. 12. The overview of the embodied grasping task. (a) demonstrates examples of language-guided grasping for different types of tasks, (b) provides an overview of human-agent-object interaction, (c) shows Google Scholar search results for topics of "Language-guided Grasping".

04. 技术路线

选择案例

业界前言技术路线：当前具身智能算法路径主要分为两条

- OpenAI 与 Figure 合作为代表的**分层决策模型**
- Google RT-2 为代表的**端到端模型**，e.g. PaLM-E

OpenAI

路线选择 I

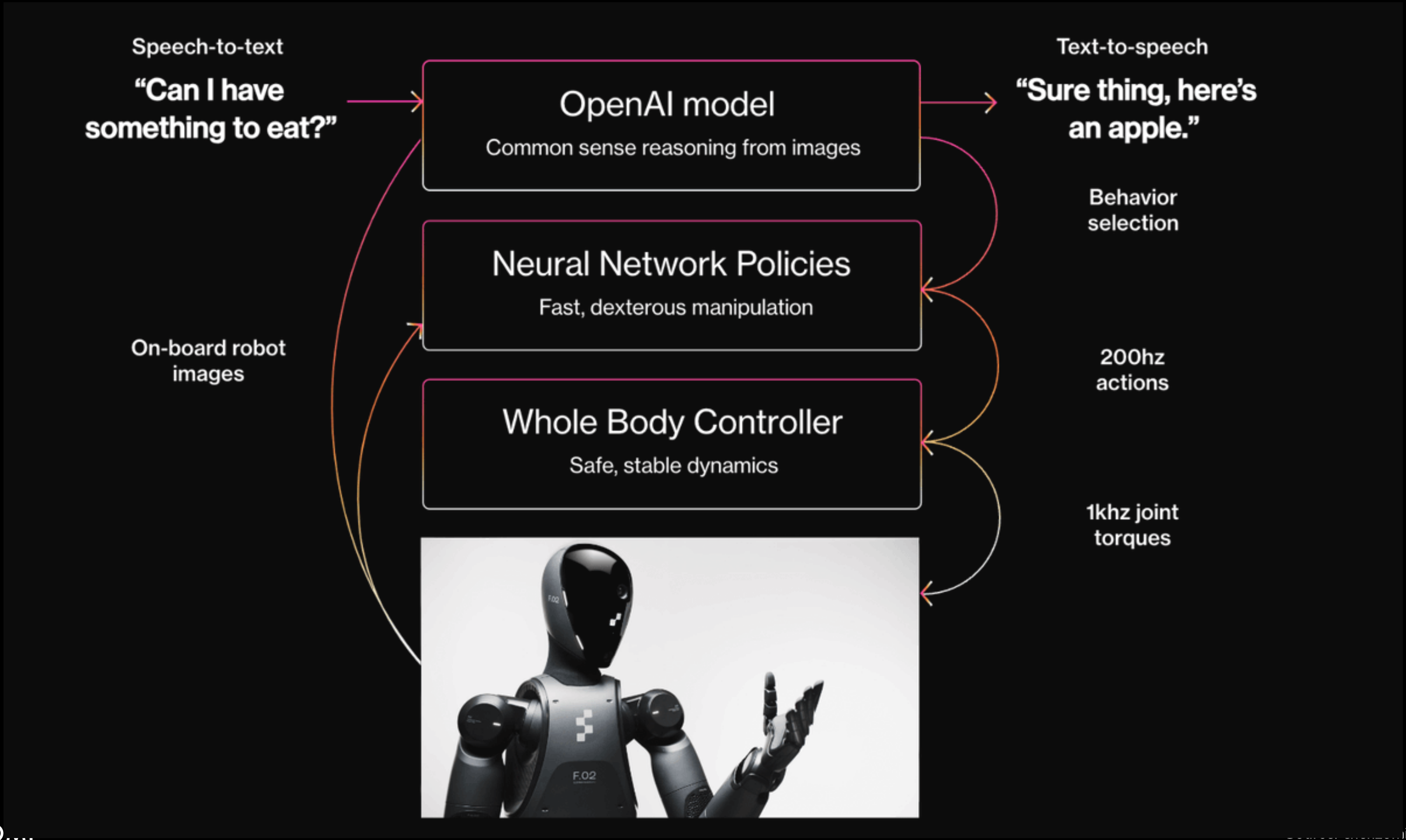

Speech-to-text
“Can I have something to eat?”
OpenAI model
Common sense reasoning from images
Text-to-speech
“Sure thing, here’s an apple.”
Behavior selection
Neural Network Policies
Fast, dexterous manipulation
On-board robot images
200hz actions
Whole Body Controller
Safe, stable dynamics
1khz joint torques
F.02

Figure 技术方案

1. high-level planer 多模态大模型感知决策同时实现，模型整合任务、环境和本体感知信息；
2. low-level policy 使用 RL 模型作为具身模型，实现从大模型的环境感知到动作的规划；
3. 最后，传统运动控制算法 whole body controller 输出机器人控制的力矩实现最终动作。

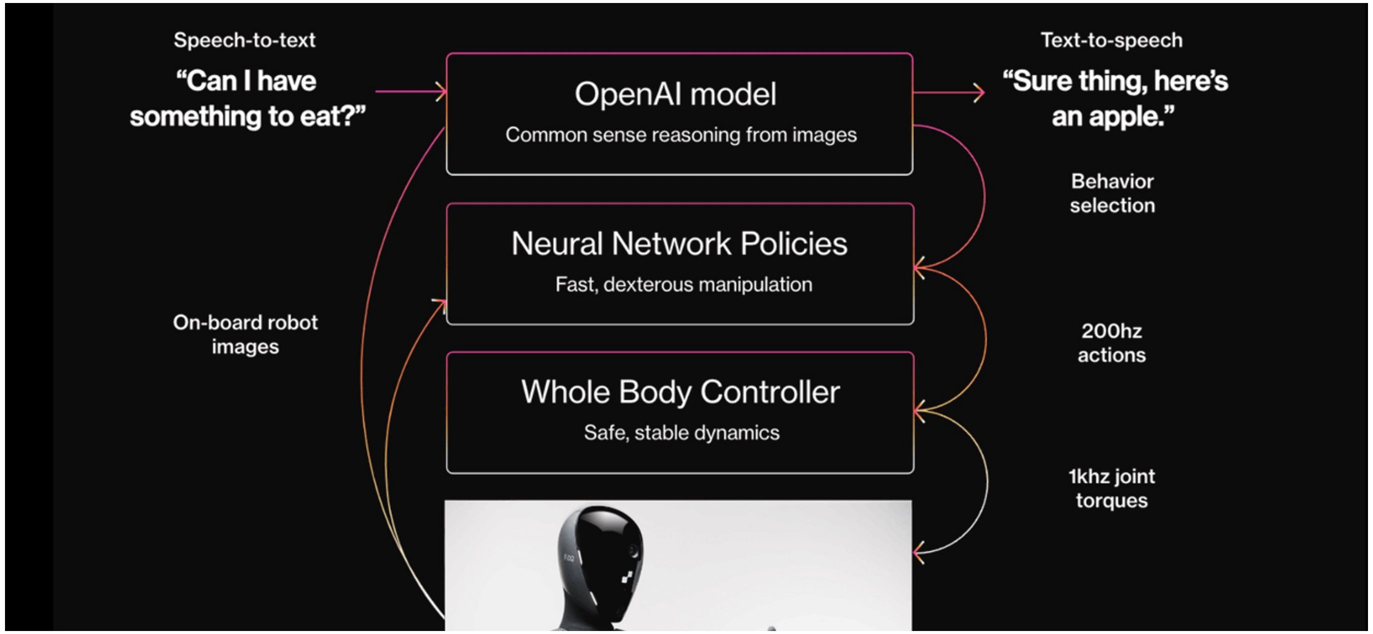


Figure 技术方案

- Figure 技术方案里面，分为 high-level internet-pretrained models + learned visuomotor policies。

优点

- 分层架构实现难度相对简单，逻辑结构清晰

缺点

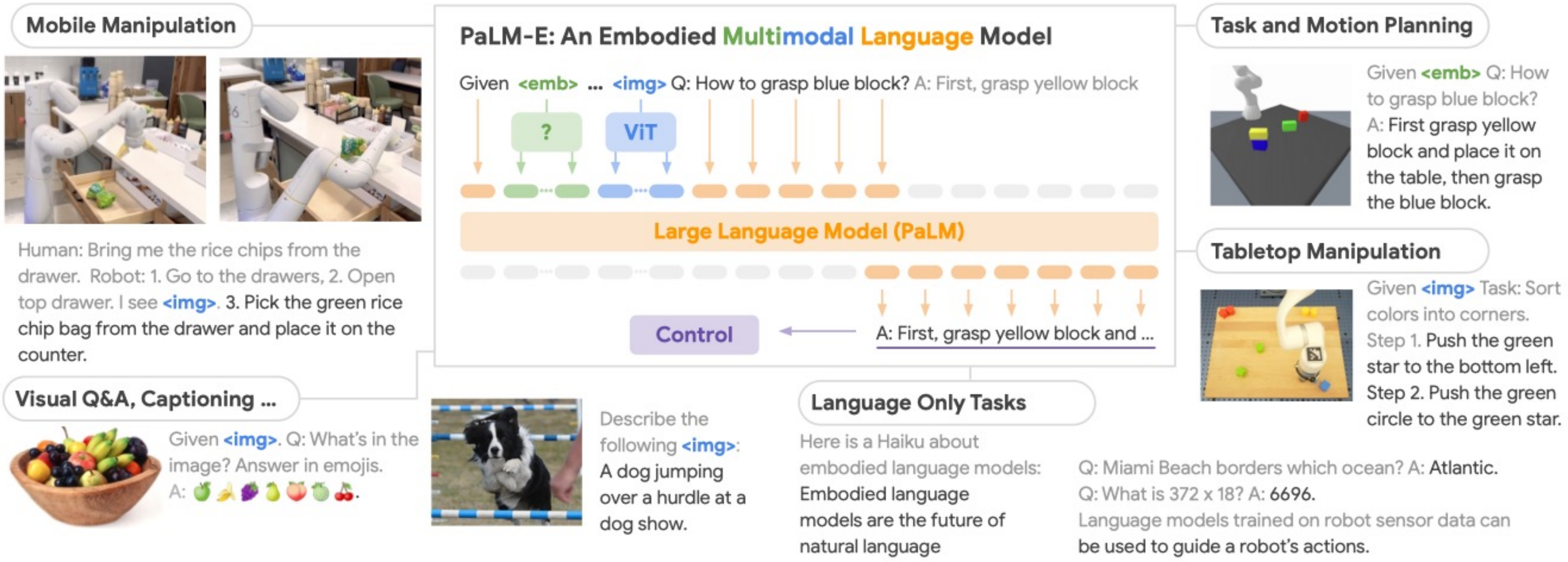
- 不同步骤间融合和一致性，是主要难点

https://palm-e.github.io/#demo

路线选择 2

PaLM-E: An Embodied Multimodal Language Model

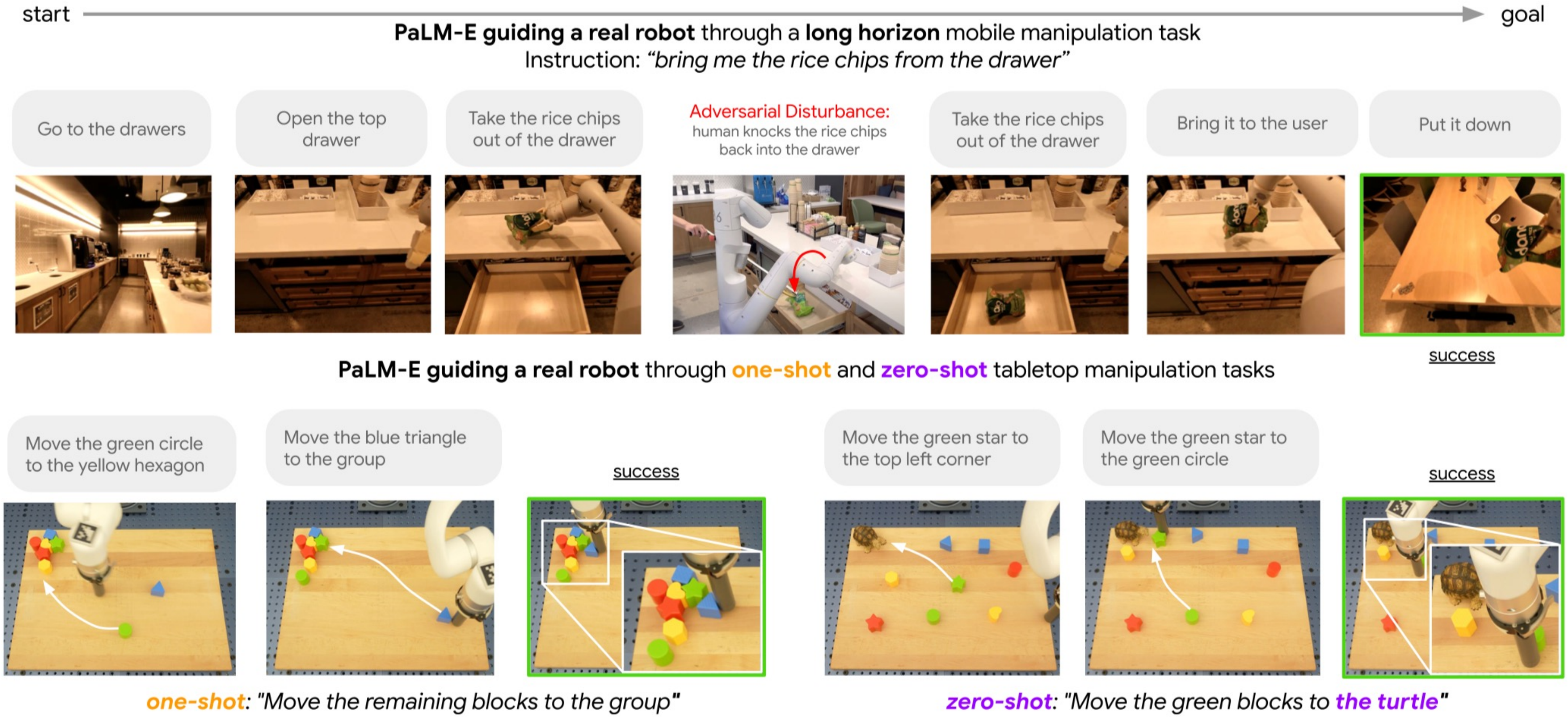
- 先在大规模互联网数据上预训练 VLMs，然后在机器人任务上微调。
- 输入是任务和对象的组合，输出是一系列动作。
- 利用大模型完成从输入到感知、推理、决策和行为指令输出的全过程。



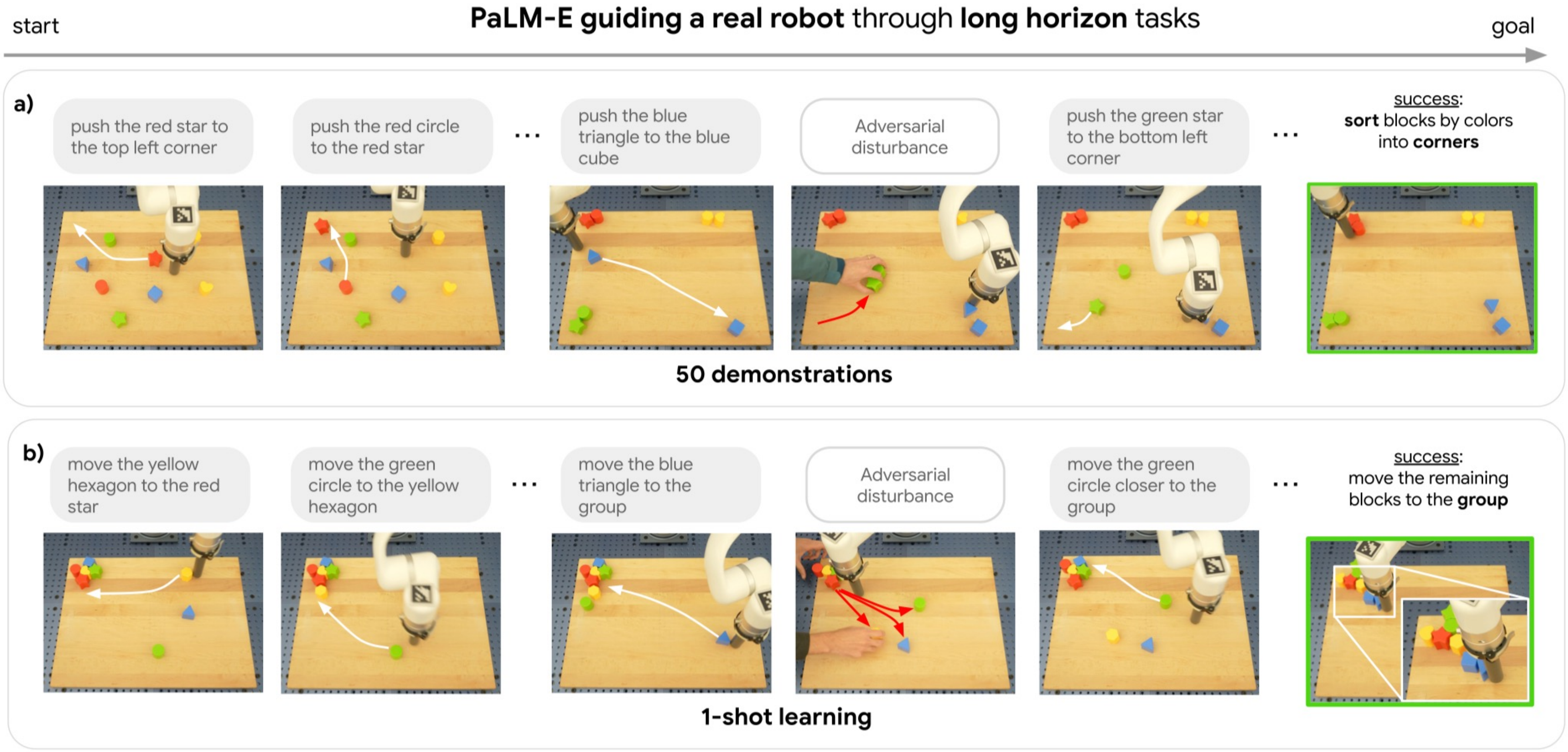
PaLM-E: An Embodied Multimodal Language Model

- 该模型的基座是之前 google 发布的预训练模型 PaLM（5620 亿），然后接上机器人，也就是具身（Embodied），所以该模型的名字为 PaLM-E（PaLM + Embodied）
- PaLM-E 通过分析来自机器人摄像头的数据来实现这一点的，整个过程不需要对场景表示进行预处理。这样一来，就不需要人类进行预处理对数据做出注释，机器人控制更加自主。

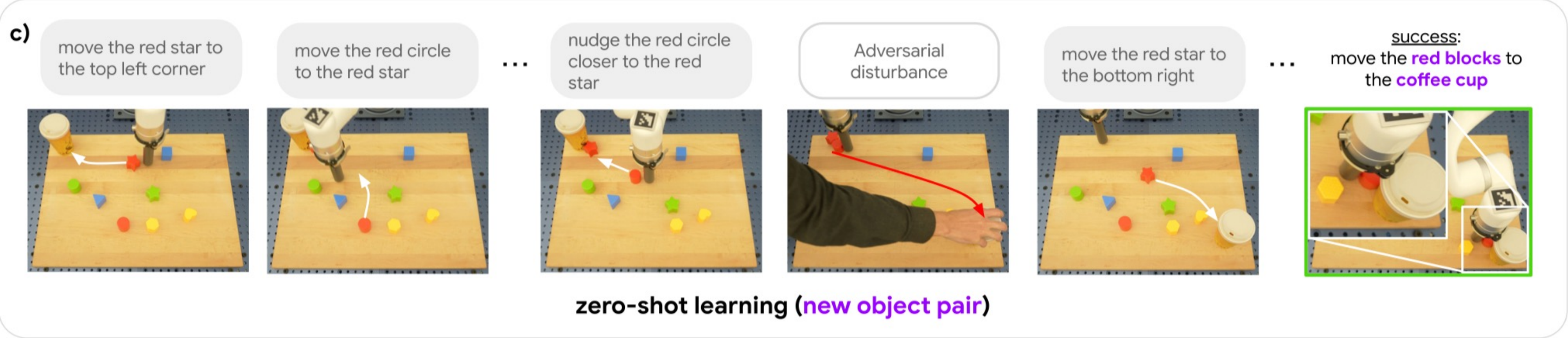
PaLM-E: An Embodied Multimodal Language Model

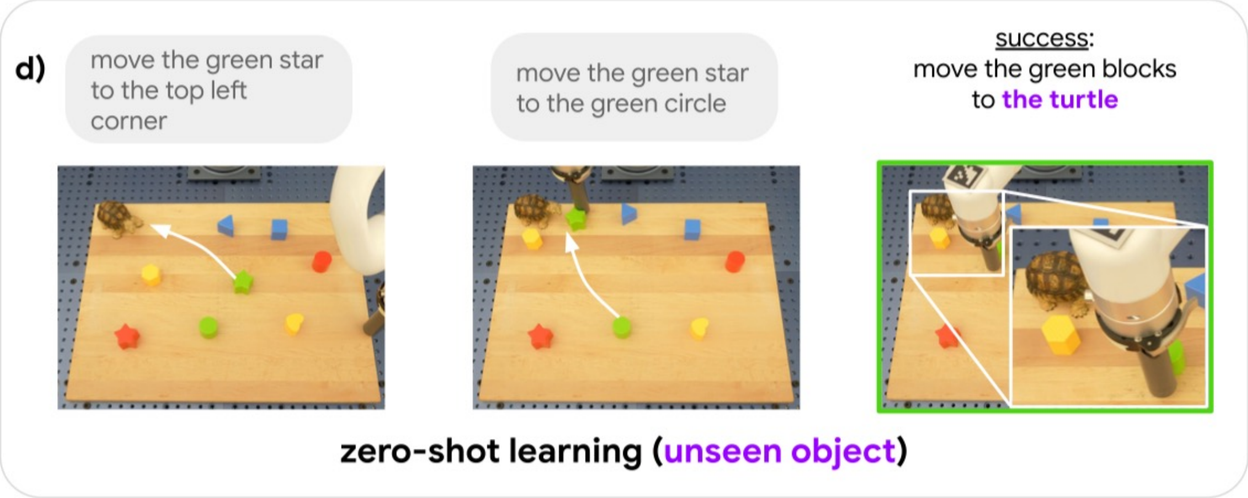


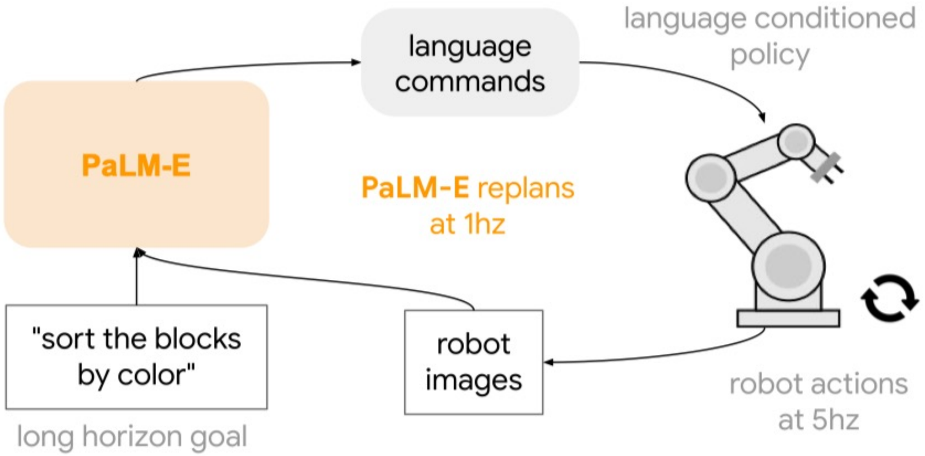
PaLM-E: An Embodied Multimodal Language Model



PaLM-E: An Embodied Multimodal Language Model







PaLM-E: An Embodied Multimodal Language Model

优点

- E2E 方案看起来更加完美，减少误差传递；
- 具身大模型观察到了能力涌现能力；
- Scaling Law 是其智能迭代一条稳定路径。

缺点

- 需要海量数据进行训练才能逐渐泛化；
- 全程调用大模型，资源消耗巨大；
- 机器人调用万亿规模大模型执行动作缓慢。

05. 思考与总结

具身智能中，硬件是一切的基础

整机伙伴	整机机器人
算法使能	数据采集 + 仿真平台 大模型 ｜ 强化学习 ｜ 工控与轨迹 ｜ 多模态
算力底座	算力芯片
机械装置	灵巧手 ｜ 支撑腿 ｜ 车轮底座
物理硬件	减速器 ｜ 轴承 ｜ 执行器 丝杠 ｜ 电机 ｜ 轮轴

数据工程问题

1. **数据稳定性：**需要自研高性能稳定硬件。如果硬件非自研，采集到的数据不适配，一旦硬件规格修改，需要重新采集数据。跨硬件算法目前还处于论文阶段。
2. **数据工程：** 涉及数据采集、数据管理、数据处理、数据组织以及与大模型和控制算法的闭环开发，需要强大数据组织能力团队。目前国内对数据研究的团队较少。

选择完算法路线的下一个难点

1. 无论何种算法方案，都需要搭建起一套完整数据收集系统，形成数据飞轮，这一套完整循环框架是当前具身智能公司的算法核心竞争力。

具身智能的估值

因估值逻辑如下，硬件、数据、算法分别都是0 ~ 1分：

1. **硬件**：如果没有自研硬件，从底层会严重受制于硬件公司；除非科研，产业落地会收到极大影响。
2. **数据**：看核心团队有没有大规模数据工程经验，数据工程经验积累尤为重要。
3. **算法**：要有顶级算法团队，即使使用开源算法，也需要顶级算法团队去消化适配。

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/AIFoundation