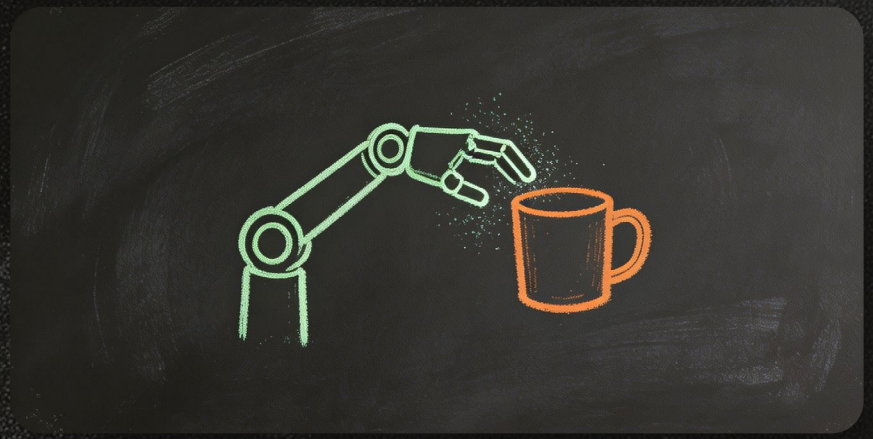
# 具身智能导论

## 从视觉抓取到模仿学习



报告人 刘孟茁

助理研究员

CIOMP, CAS

# 三个核心问题贯穿全文

01

What

什么是具身智能?

02

Why

为什么需要具身智能?

03

How

如何实现具身智能?

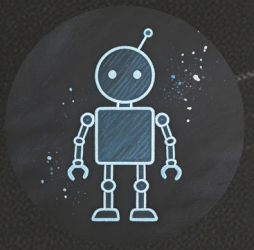
# 今天我们学什么？

COURSE AGENDA

## 01

### 离身智能

理解智能的起源与发展基础



## 02

### 具身智能

探索智能体与环境的交互机制

## 03

### 核心技术体系

感知、决策与控制的技术融合

## 04

### 大模型驱动

基于LLM的具身智能新范式

## 05

### ACT算法

经典的具身策略与规划算法

## 06

### 总结与未来展望

技术落地挑战与未来发展方向

# 第一部分：离身智能 / Disembodied AI (DAI)

💡 核心认知：理解「离身智能」是深入探究「具身智能」的逻辑前提与理论基石

## 🧠 “传统”AI主流：离身智能体系

流派 A：符号主义 (Symbolism)

基于逻辑推理与规则符号的自上而下认知

流派 B：联结主义 (Connectionism)

基于神经网络与统计学习的自下而上拟合

✅ 符号主义 + 联结主义 ≈ 离身智能

## 🚪 转向的契机：行为主义 (Behaviorism)

不同于传统的静态数据推演，行为主义强调智能体与环境的动态交互。它打破了封闭的符号系统，为AI从“离身”走向“具身”提供了关键的理论支撑。

通向：具身智能研究

Embodied AI Research

# 符号主义 — 强调"表示"

## 💡 核心思想

将人类积累的知识与经验抽象表示为规则与符号，通过严密的逻辑推理链条来模拟人类思维，从而解决特定领域的复杂问题。

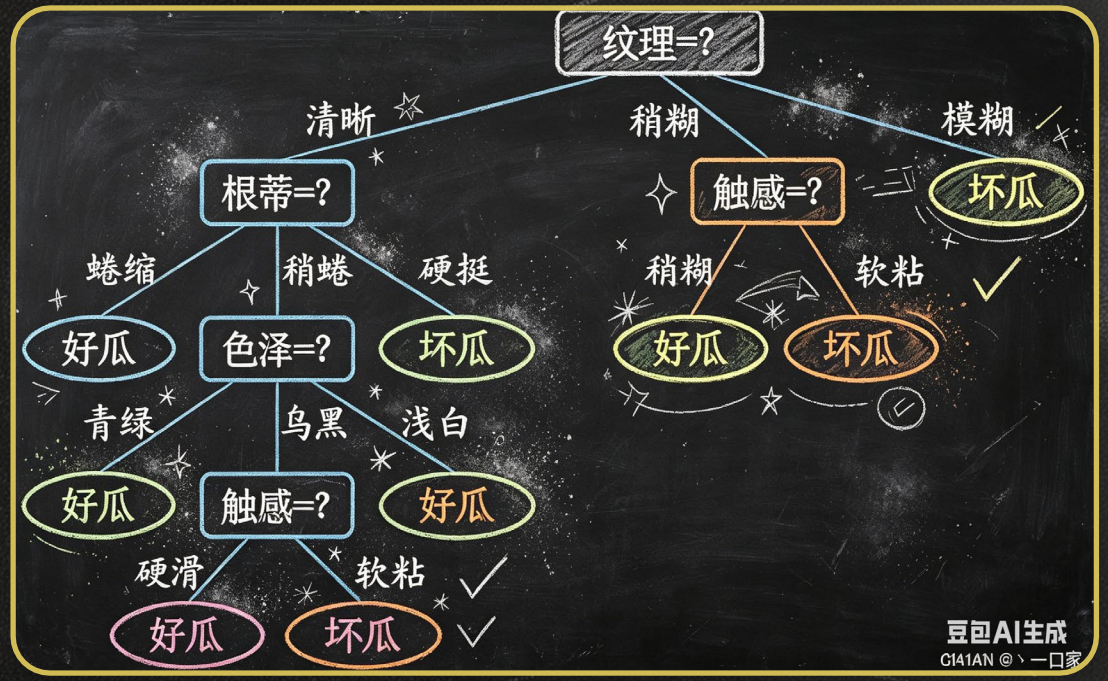
## 🛠️ 代表方法

经典的算法与系统包括：决策树、专家系统、逻辑编程（如Prolog）以及基于规则的产生式系统。

## 📊 核心优劣势

✅ 优势：逻辑推导过程完全透明，具备极强的可解释性，知识表达直观清晰。

⚠️ 局限：难以处理现实世界中广泛存在的不确定性、模糊性，以及海量的非结构化数据。




符号主义典型模型 — 决策树结构

# 联结主义 — 强调“计算”

## 💡 核心思想

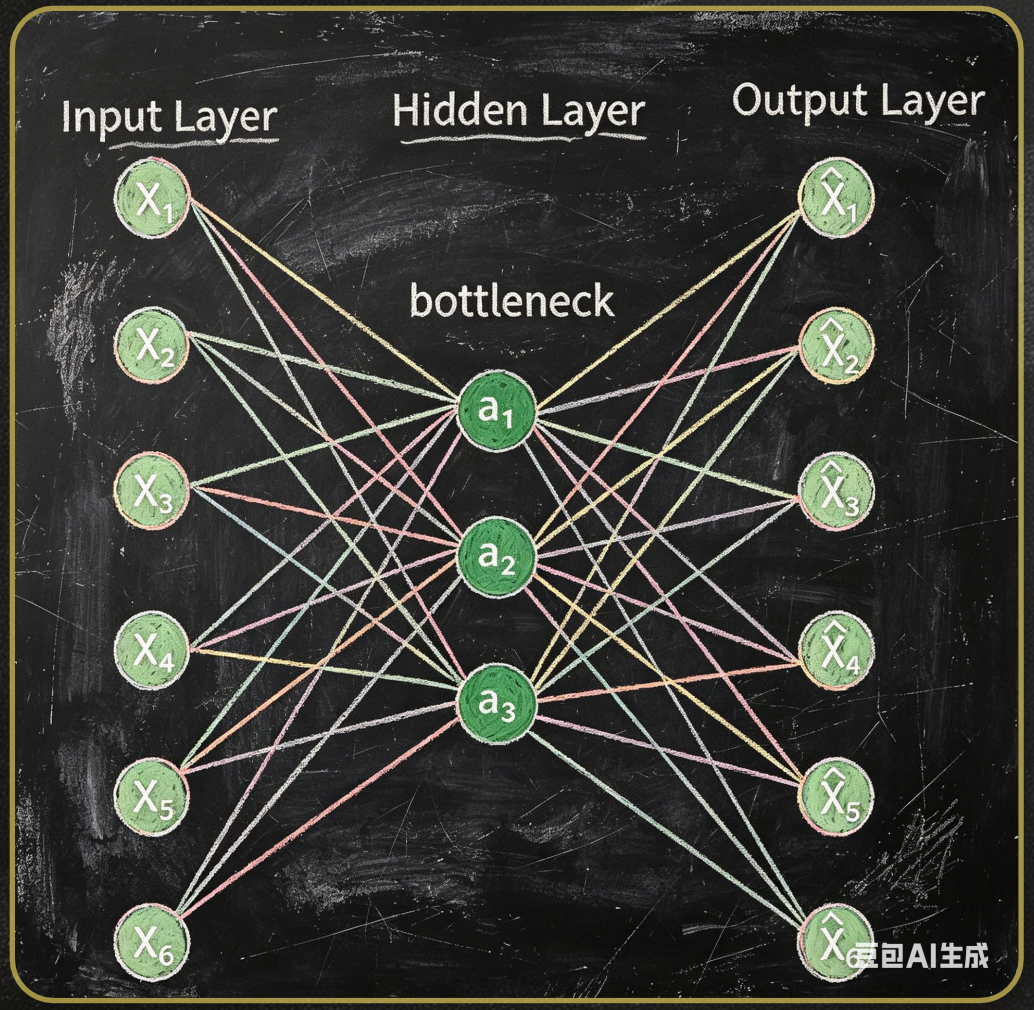
模仿大脑神经网络的信息传递机制，通过向模型输入海量数据进行训练，使其自动从数据中学习并总结经验规律。

## 🧠 代表方法

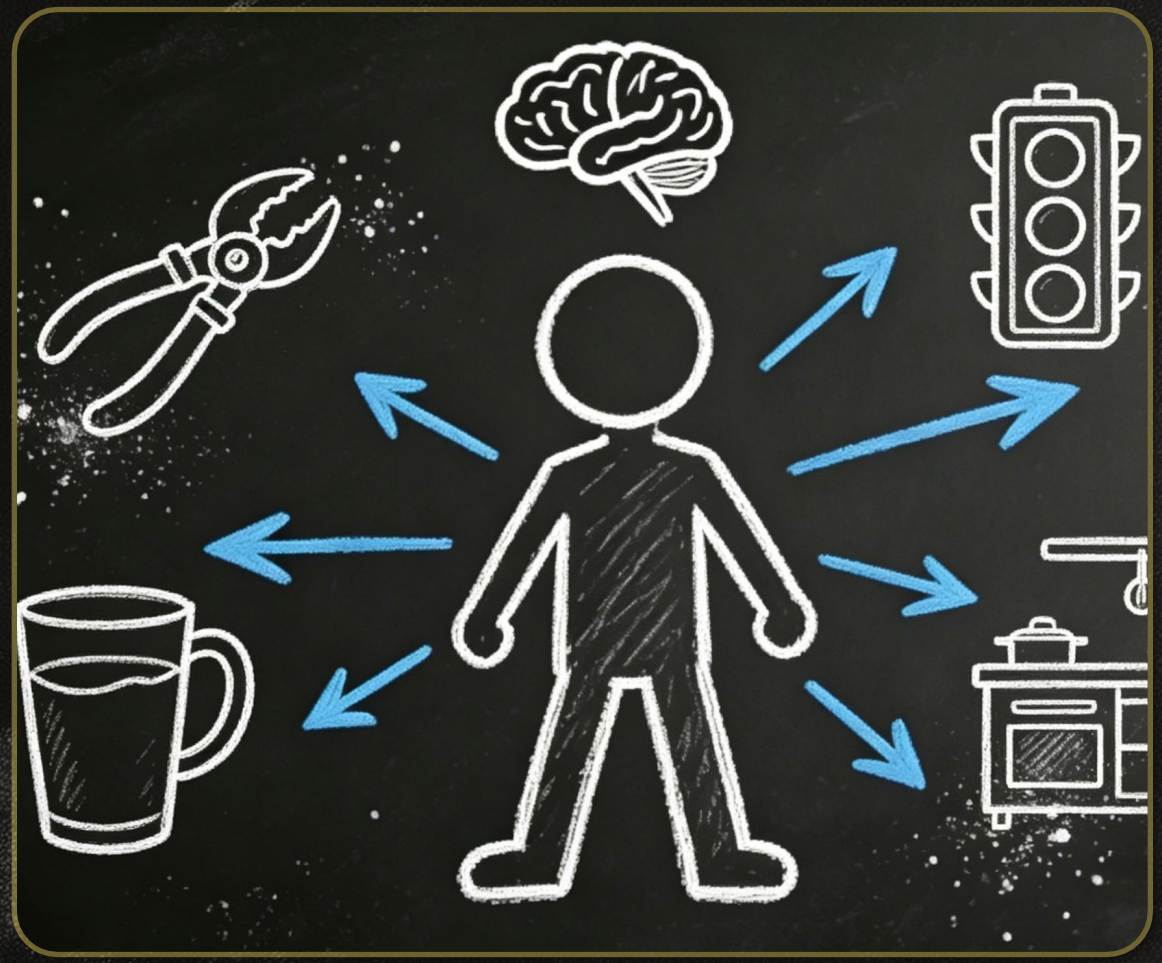
以深度学习架构为核心，典型代表包括CNN (卷积神经网络)、RNN、Transformer以及当下热门的大语言模型 (LLM)。

## ⚖️ 优势与局限

✅ 优势：自动特征提取能力强，泛化性能优异，擅长处理图像、语音等非结构化数据。
❎ 局限：高度依赖海量标注数据，可解释性弱（黑箱问题），缺乏对物理世界的交互逻辑。



# 行为主义 — 强调“交互”



## 💡 核心思想

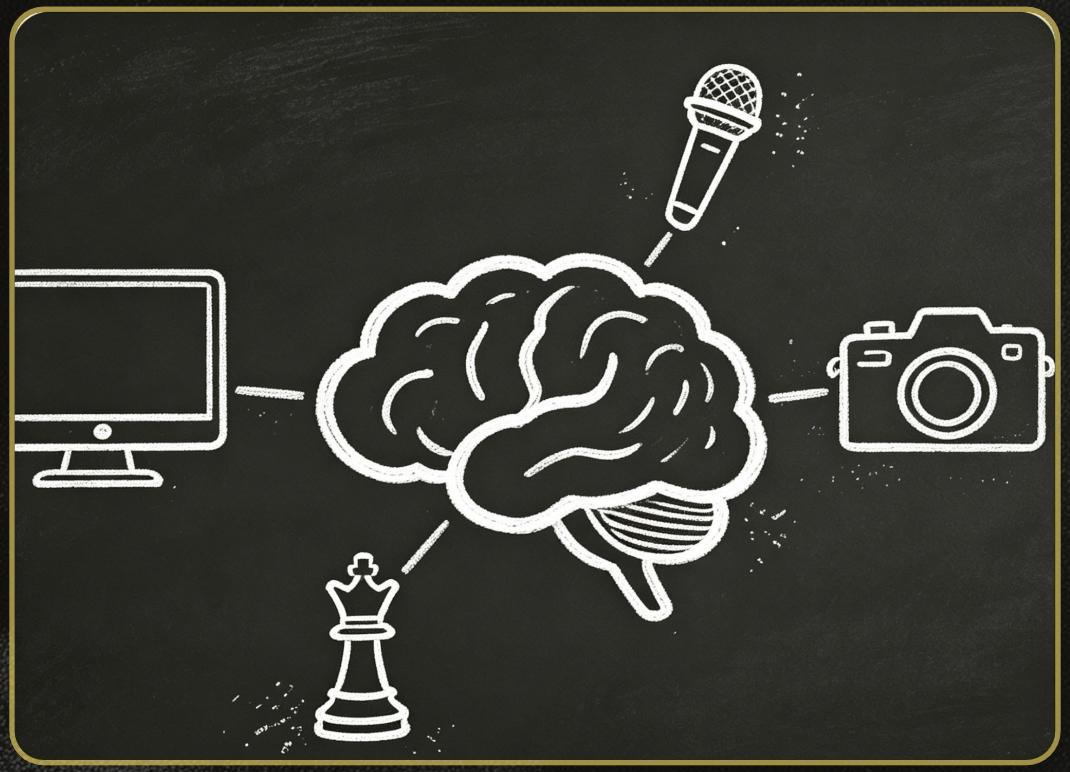
智能并非孤立存在于大脑，而是由脑、身体与环境三者协同作用、动态交互产生的。

## 🔑 关键机制

在身体与物理环境的持续互动中，通过主动的“感知”与“物理操作”，智能才得以涌现和发展。

⭐ 这直接指向了 —— 具身智能！

# 离身智能的核心特征



## 无实体依赖的软件形态

完全以纯软件形式存在于数字空间，不依赖任何物理身体作为运行载体。

## 逻辑符号与海量数据驱动

基于严密的符号处理与逻辑推理，并结合海量数据的训练来不断优化模型。

## 信息的抽象算法处理

剥离了具体的物理感知，其本质是对输入信息进行纯粹、抽象的数学算法运算。

## 成熟的典型应用领域

广泛应用于计算机视觉(CV)、自然语言处理(NLP)以及AlphaGo等经典棋类AI中。

# 离身智能的局限性

## 认知智能的「天花板」

认知智能无法靠离身方法解释。
人类的高级认知高度依赖具身经验与环境的持续交互。

仅靠符号表示和逻辑计算无法覆盖全部任务。
面对非结构化的现实世界，传统计算显得力不从心。

*传统AI困境：脱离了「身体」的智能是不完整的。*

### Pick-and-Place 抓取实验

人类无需精确计算物体的三维坐标就能灵活抓取，而传统算法却需要海量的参数运算才能勉强完成。

💡 结论：但这还不够！我们需要具身智能

# 传统 Pick-and-Place 的方法

01

RGBD相机检测

利用深度相机
获取场景点云数据

02

CV识别物体

基于视觉算法
定位并识别目标

03

计算抓取点位

规划最优的
机械臂抓取坐标

04

点云确定距离

分析三维点云
计算目标空间距离

05

IK计算关节

通过逆运动学
求解各关节角度

06

控制机械臂抓取

下发指令至伺服
执行最终抓取动作

# 人类是怎么做的？我们从来不计算坐标

## ❌ 传统的「计算」思路

在传统的机器人控制中，我们试图模拟“计算”。每一个动作都需要精确计算关节坐标、解算复杂的逆运动学(IK)方程，依赖大量的数学推导与预设参数。

$$\theta = \arccos((p^2 + a^2 - b^2) / 2pa) + \Sigma(...)$$

我们的大脑里，根本没有这些！

这种“纯计算”的方式繁琐、僵化，无法适应动态变化的真实物理环境。

## ✅ 人类的「具身」思路 (智能)

直接通过视觉感知目标，大脑将视觉信息**直接映射**到肢体动作，无需中间的复杂数学推导。

💡 这才是更直接、更泛化的智能方式

# 儿童实验 — 认知科学的启示

## Baldwin 实验

**01 自由把玩**

幼儿自主探索并把玩
不熟悉的新奇物品

**02 视觉隐藏**

将物品完全藏入
不透明的容器中

**03 语言输入**

实验者对着容器
说出一个新词 “dax”

**04 建立连接**

幼儿能准确将 “dax”
对应到之前把玩的物品


### 💡 认知科学核心启示

儿童并非单纯依靠“看”来学习语言，而是通过「位置-空间-物体」的物理交互经验，建立起对抽象新词的理解。

# 儿童为什么能做到?

## 核心逻辑 · Core

Linking objects to locations using attention.

利用注意力将物体与空间位置建立强关联

## 物理世界 · Rules

A real object distinguished by unique location.

真实世界中，物体由其唯一的物理位置所定义

## 认知闭环 · Conclusion

儿童通过与物理环境交互，建立起完整的认知链条：

位置 → 空间 → 物体 → 名称

# 从离身到具身 — 智能的新视角

## 离身智能

Disembodied Intelligence

侧重模拟大脑的纯理性逻辑活动，将智能视为独立的符号处理系统。

身体仅作为被动的“执行器”，不参与智能的生成过程，与环境隔离。

## 具身智能

Embodied Intelligence

智能产生于“身体”与“环境”的持续交互中，认知根植于物理体验。

强调“感知-执行-反馈”的完整闭环，主动适应环境变化，具备涌现性。

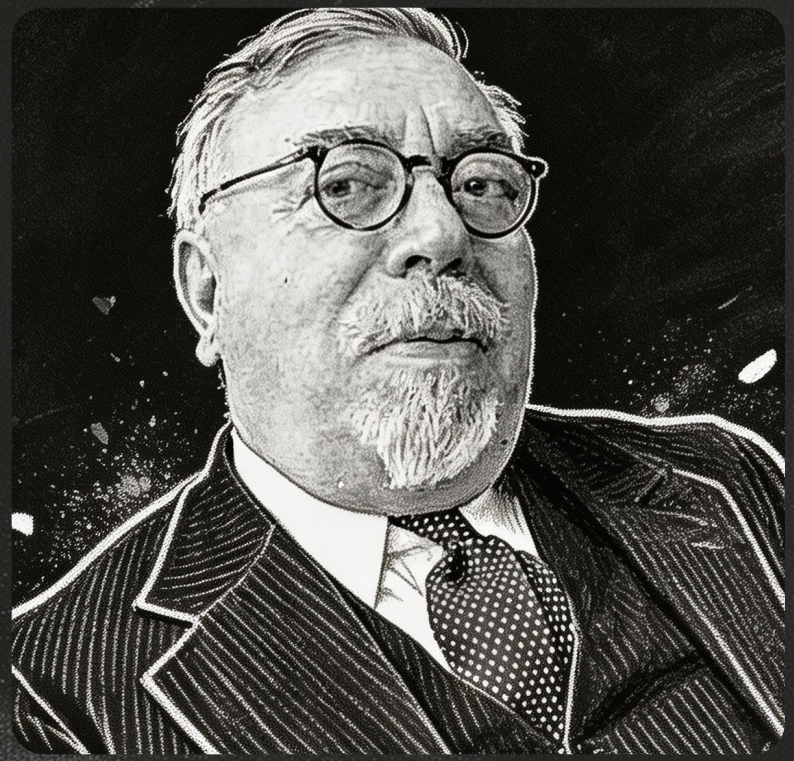
# 02

# 第二部分 / 具身智能的历史与内涵

HISTORY · DEFINITION · CONSENSUS

历史溯源 · 概念定义 · 业界共识

# 具身智能的思想溯源



诺伯特·维纳

Norbert Wiener | 控制论之父

## 先驱人物

20世纪著名数学家，在电子工程、计算机科学、人工智能等领域贡献卓越。

## 核心著作 · 《控制论》

首次打破了生物与机器的界限，将反馈机制作为核心概念。

## 关键时间节点

1935-36年：任清华客座教授，提出“维纳滤波器”；1948年：《控制论》正式出版，标志学科诞生。

## 思想的深远影响

为后续的信息论、图灵机模型、冯·诺依曼架构奠定了坚实的理论基础。

# 为什么具身智能沉寂了20年?

## 1980s · 主流确立

AI 研究的核心方向确立为“离身智能”，学术界与工业界的焦点集中在计算机视觉 (CV)、自然语言处理 (NLP) 以及棋类博弈等纯算法任务上。

## 2000s - 2010s

行业冷门

具身智能研究更多局限于机器人学领域，成果主要发表在 IROS、ICRA 等专业会议，与 IJCAI、NeurIPS 等 AI 主流顶会脱节，处于边缘发展状态。

# 具身智能的复兴 — AGI 共识


## 💡 背景认知

随着技术的快速演进，整个社会对「如何实现通用人工智能 (AGI)」已经形成了普遍的行业共识，AGI 不再仅仅是遥远的科幻设想。

## 🔑 核心关键

大语言模型 (LLM) 的爆发让人们切实看到了AGI的可能性；而具身智能是让智能体真正落地物理世界、实现完整AGI的必要核心要素。

## ✨ 最终结论

LLM + 具身智能 = 实现 AGI 的「缺一不可」

# 具身智能的定义 — 两类视角

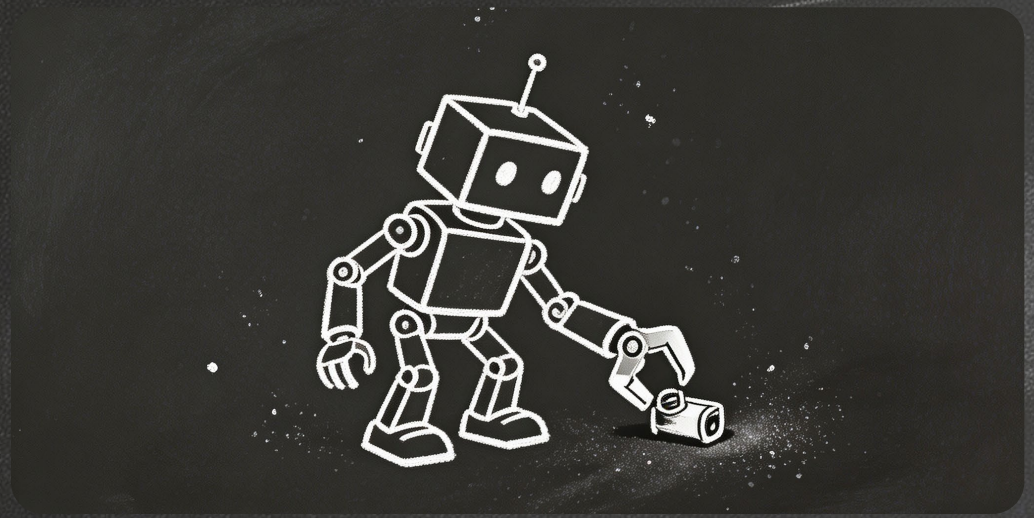

## 理论视角 · 认知科学

具身智能在身体与环境相互作用中，通过信息感知与物理操作过程可以连续、动态地产生智能

------ 刘华平

强调具身交互对智能的深层塑造。身体在与物理世界的持续互动中，产生并动态影响着智能的形成与发展。

*视角来源：神经科学 / 认知科学研究*



## 应用视角 · 机器人实践

具身的含义在于与环境交互以及在环境中做事的整体需求和功能

------ 李飞飞

核心目标是赋予AI物理实体。智能体通过拥有的物理身体，与真实环境进行自主交互，从而感知信息并完成具体任务。

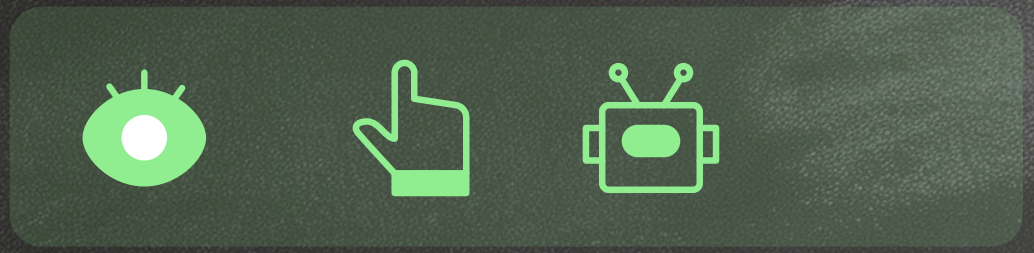
*视角来源：机器人工程 / 人工智能应用落地*

# 业界普遍共识

## ❌ 思维惯性

将智能仅局限于数据的抽象处理与数学问题的求解。误认为只要拥有强大的 LLM 或 CV 算法，就实现了完整的“智能”。

## ✅ 业界共识



真正的智能必须具备具身属性：对周围物理环境的实时感知、对复杂场景的深度理解，以及在现实世界中的自主行动与操作能力。

💡 核心观点：LLM 与 CV 是关键技术底座，但并不等同于“具身智能”

# 03

# 第三部分 / 具身智能的核心技术体系

感知 · 决策 · 行动 · 反馈

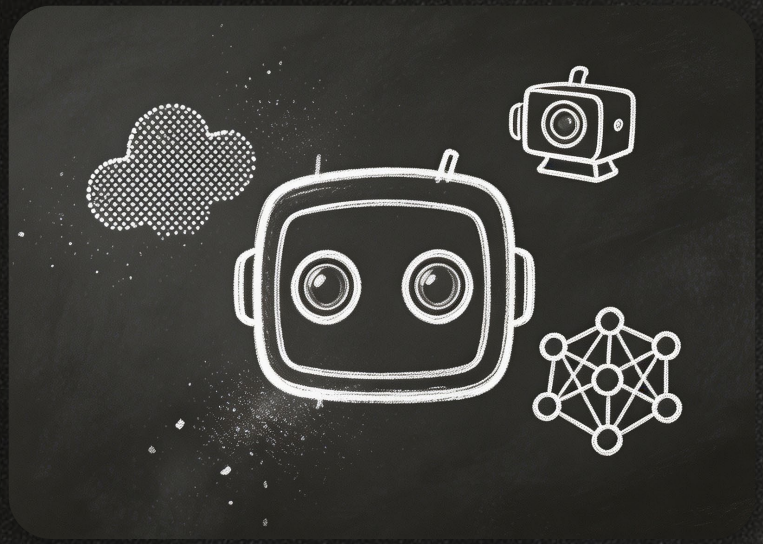
感知

决策

行动

反馈

# 感知 — 采集与理解外部世界



## 核心作用

对外部环境信息进行实时采集与处理，建立精准的环境感知模型，为后续的决策规划模块提供坚实的数据支撑。

## 关键技术栈

结合深度相机与激光雷达数据，运用**3D重建**技术还原三维空间，利用**VSLAM**（视觉SLAM）实现动态环境中的定位与地图构建。

## 智能视觉基础

依托**大模型**的海量数据理解能力，结合**多模态大模型**对图像、语音等异构信息的融合分析，实现更高级的语义感知与理解。

# 决策 — 任务规划与推理

## 核心作用

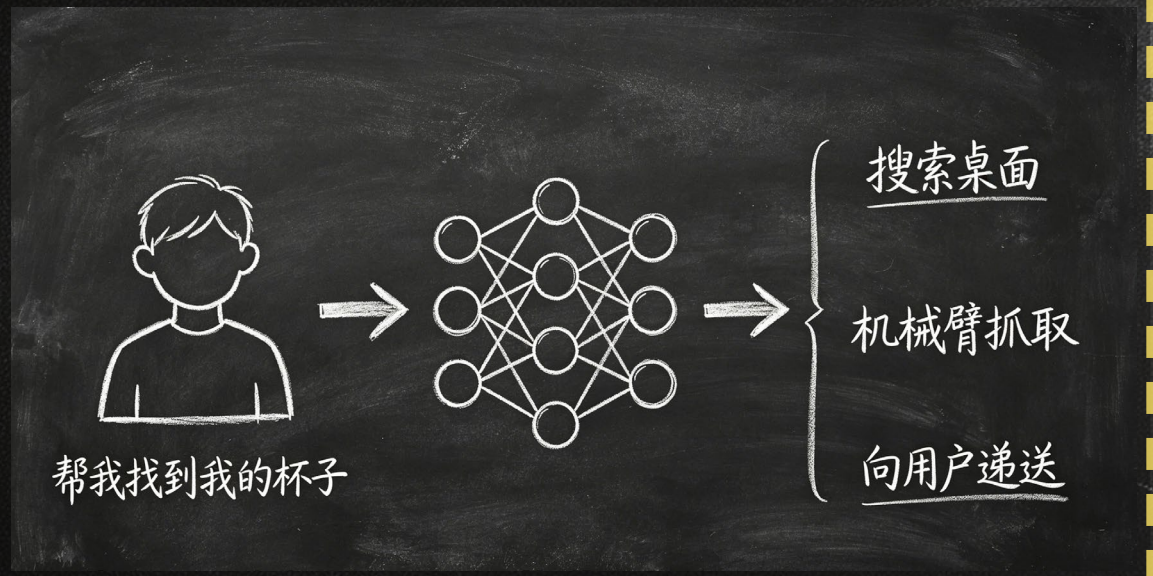
作为机器人的“大脑”，根据感知信息与任务目标进行综合规划与逻辑推理，最终生成可执行的具体决策指令。

## 关键技术路径

- 传统：人工编辑逻辑规则 / 强化学习(RL)
- 前沿：LLM理解任务生成代码 / VLA视觉语言规划

## 决策流程示例

输入“找到我的杯子” → 拆解：[搜索桌面] → [机械臂抓取] → [向用户递送]




*图示：典型的决策逻辑流向示意*

# 行动 — 执行具体动作

## LOCOMOTION

**移动与导航**

自主路径规划 · 实时避障 · 定点导航

## MANIPULATION

**双手精细操作**

物体精准抓取 · 工具灵活使用 · 双臂协同

## INTERACTION

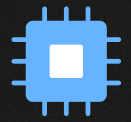
**环境智能互动**

场景动态感知 · 意图理解 · 自适应响应

## —— 核心实现路径 ——

经典控制方法

PID控制 / 运动学逆解



大模型智能驱动

LLM规划 / 多模态理解

混合增强架构

高层规划 + 底层控制闭环

# 反馈 — 持续优化与调整

## ⚠ 无反馈的困境

长程任务若缺失反馈机制，系统将无法感知执行偏差，极易沿着错误的路径持续运行，最终导致任务彻底失败。

## ♻ 反馈的核心价值

- 实时修正，大幅提升环境感知的精度
- 依据偏差数据，即时调整决策与执行策略
- 持续优化动作细节，形成任务闭环

# 感知-决策-行动-反馈 — 完整闭环

## 核心机制

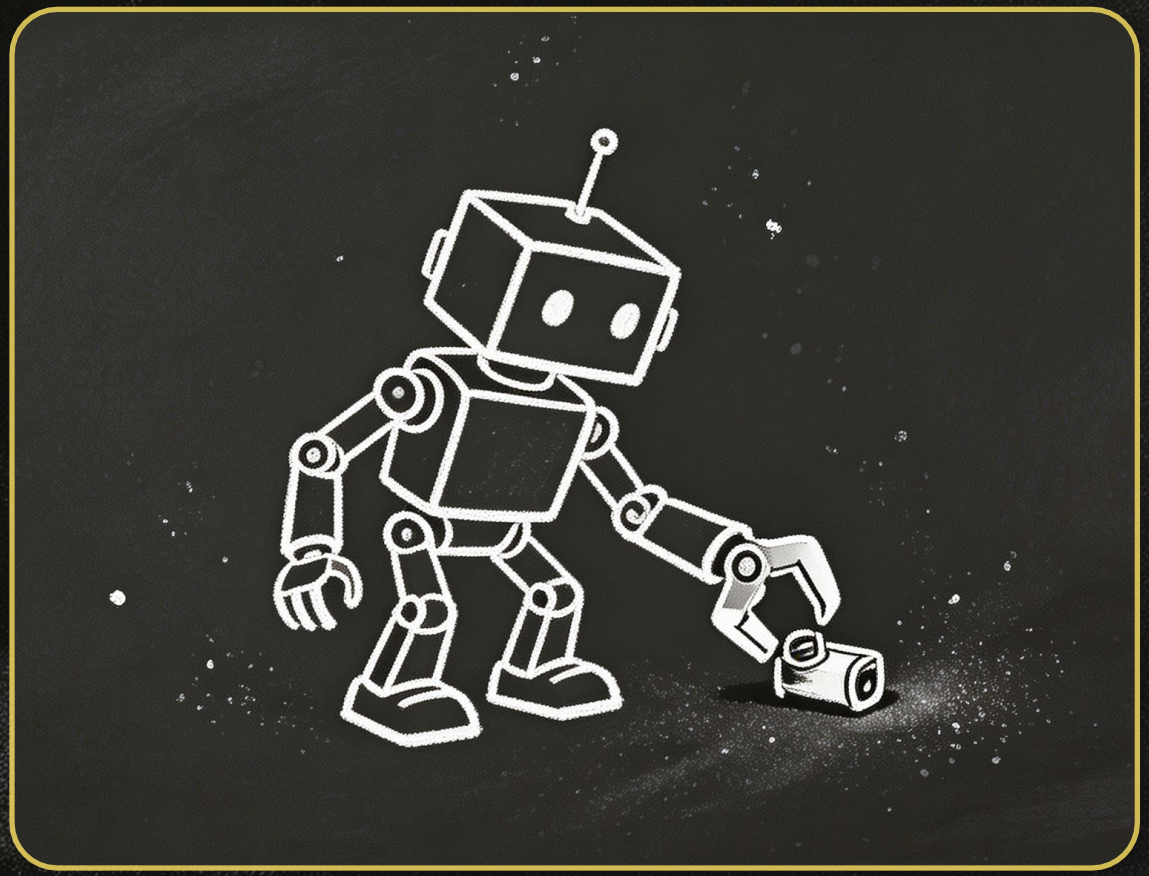
四大模块紧密耦合，形成一个持续迭代的完整闭环，共同驱动系统实现对环境的理解与适应，最终达成具身智能的核心目标。

## 闭环作用拆解

- 感知：作为系统的“眼睛”，从环境中获取多维信息与数据输入。
- 决策：作为系统的“大脑”，基于信息拆解任务，规划最优执行路径。
- 行动：作为系统的“手脚”，精准执行规划指令，与物理世界交互。
- 反馈：作为系统的“修正器”，对比预期与结果，实时修正执行偏差。

# 案例：Go find and bring my cup

—— 用一个具体场景串起机器人的四个核心模块 ——

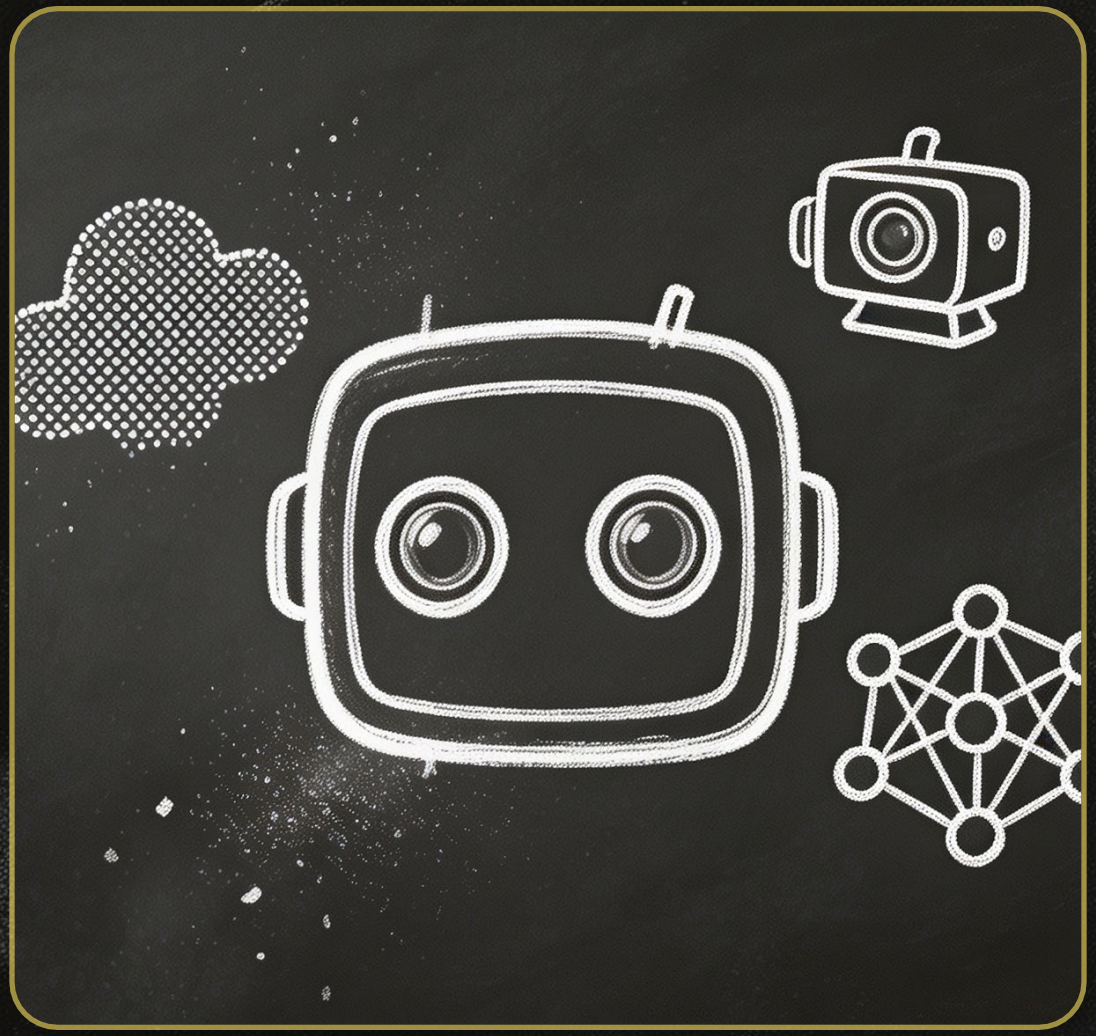


## TASK / 核心任务

“找到我的水杯，并把它递过来”

这是一个典型的轮臂机器人自然语言交互场景。这句简单的指令，背后完整串联了机器人的感知、决策、行动、反馈四大核心处理模块。

# Step 1 — 多模态感知



## 核心感知传感器阵列

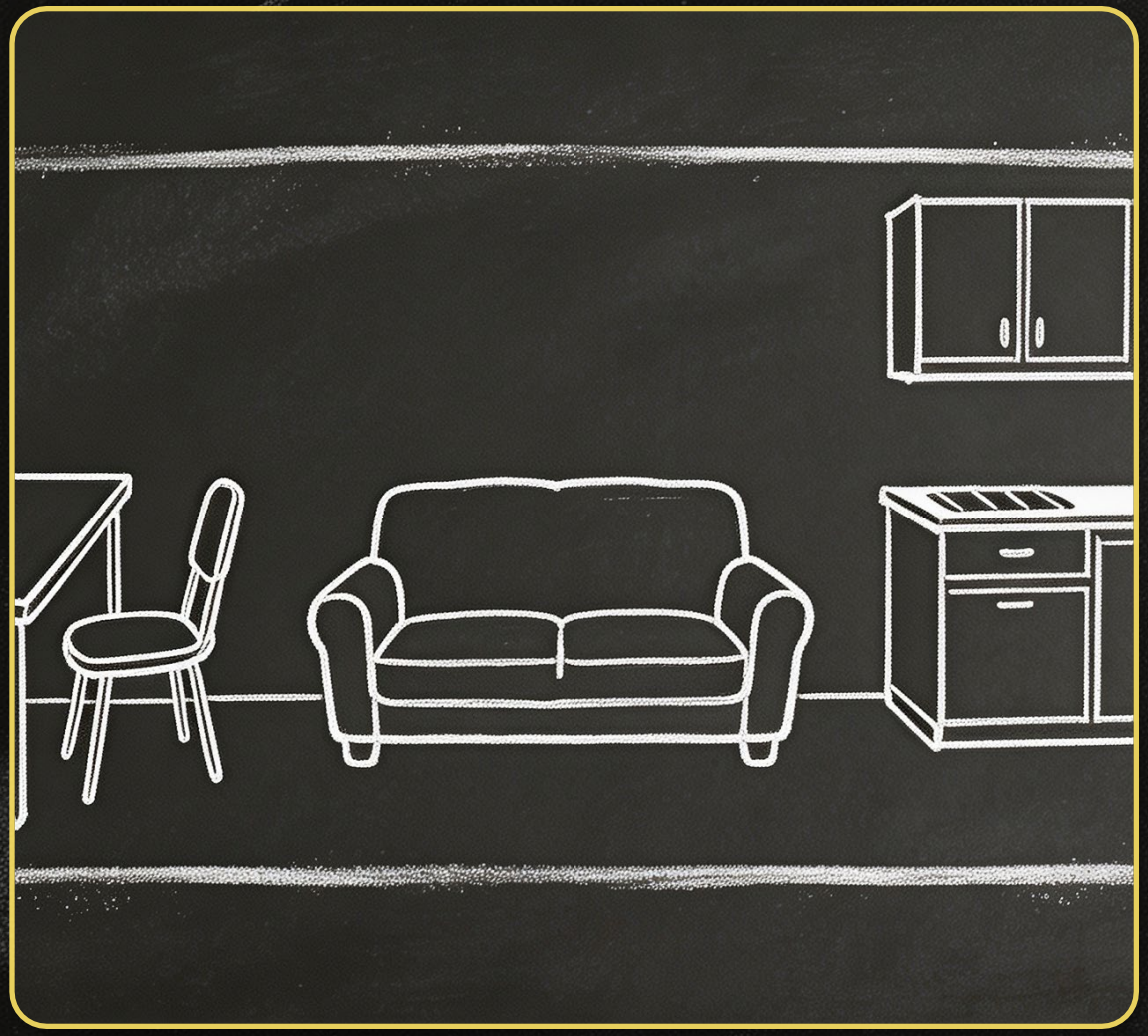
- 视觉：RGBD 相机 / 双目深度相机
- 测距：激光雷达 (LiDAR) / 超声波距离传感器
- 交互：高精度触觉反馈传感器

## 实时环境分析结果

对当前视野（3m x 3m）区域进行全方位扫描分析后得出：

“当前区域内未检测到目标物体（水杯）”

# Step 2 — 决策：目标缺失后的推理



## 搜索优先级 (Priority)

01 书桌抽屉 (Desk Drawer) — 高频使用区

02 沙发旁茶几 (Coffee Table) — 休闲停留区

## 动态搜索路径 (Strategy)

Start 书桌区域 ▶ Next 沙发区域 ▶ End 厨房区域

# Step 3 — 行动：逐一搜索

## 01 / 第一站

### 书桌抽屉

打开抽屉仔细翻找，未检测到目标水杯，继续前往下一站。

## 02 / 第二站

### 沙发边几

扫描边几置物区及缝隙，仍未发现水杯踪迹，持续搜索中。

## 03 / 第三站

### 厨房台面

感知模块成功检测到目标！立即执行抓取与递送动作，任务圆满完成。

# Step 4 — 反馈：实时修正

## 扫描无目标

当抽屉内未检测到目标物品时，系统触发负反馈，自动切换至下一预设搜索位置，确保任务不中断。

## 抓取保稳定

末端力传感器实时回传压力数据，动态调整机械臂的抓取力度，既保证物体不滑落，又防止力度过大损坏物品。

## 调整递送路径

视觉传感器持续监测环境与目标位置，一旦发现路径偏差或障碍物，立即反馈至决策层，实时修正运动轨迹。

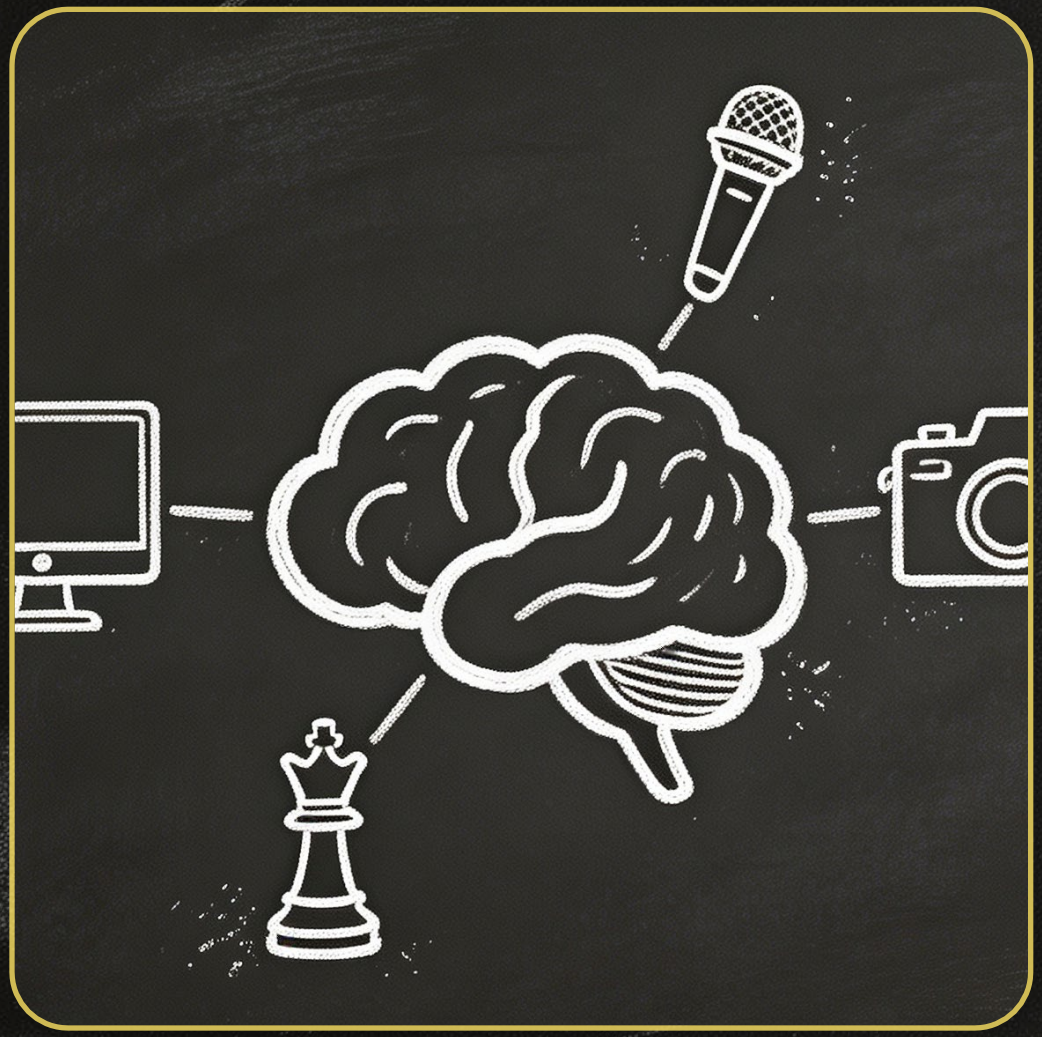
核心闭环：在任务执行的全流程中，形成**感知 → 决策 → 执行 → 反馈**的完整闭环系统，通过持续的实时修正，确保复杂任务的最终成功。

# 04

# 第四部分 / 大模型驱动的具身智能

VLA · VLN · 大模型在各模块的应用

# LLM 与具身智能的关系



## 技术基石：NLP 领域的核心方法

LLM (大语言模型) 本质上是自然语言处理领域发展出的一种先进技术路径，擅长处理文本信息。

## 驱动方式：具身智能的路径之一

基于 LLM 驱动的 Agent 架构，只是实现具身智能的众多技术方法中的一种，并非唯一解。

## 认知误区：VLA 并非全部

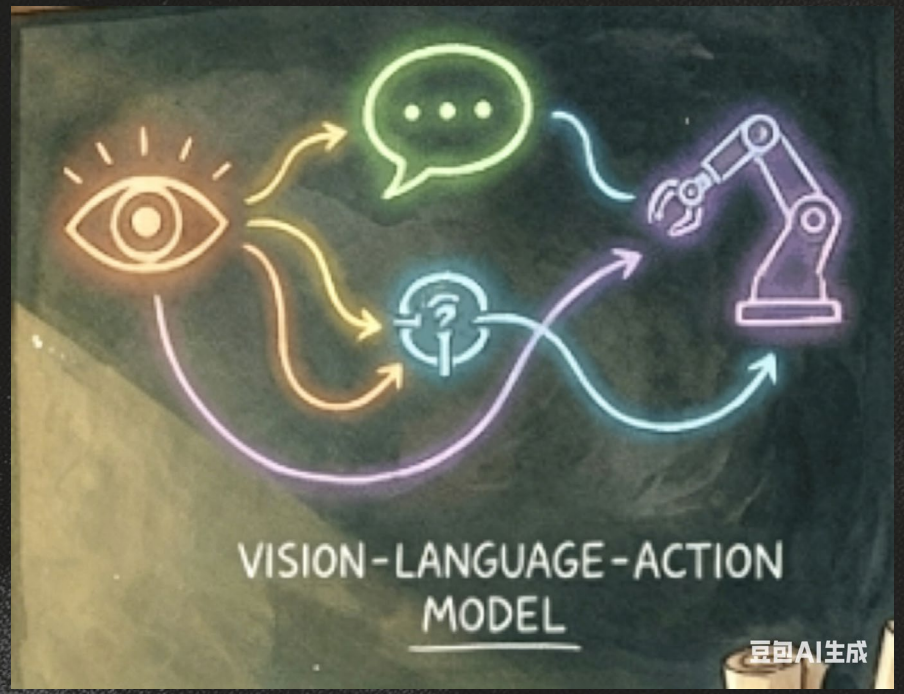
VLA (视觉语言模型) 虽然强大，但它仅仅是感知层面的融合，不能完全代表具身智能的完整闭环。

## 能力泛化：赋能四大核心模块

大模型的能力可以灵活应用于具身智能系统的感知、决策、行动、反馈这四大核心功能模块中。

# VLA — 视觉-语言-动作模型

Vision-Language-Action Model




## 核心思想

打破传统分步感知与规划的界限，实现端到端的直接映射：从视觉信息与语言指令直接生成动作。

## 模型输入

1.视觉图像：摄像头捕捉的环境画面
2.语言指令：用户发出的自然语言任务描述

## 模型输出

生成可供机器人执行的离散动作序列或连续的控制指令，直接驱动机械臂或移动底座完成任务。

## 人类类比

非常类似于人类的行为模式：眼睛(视觉)看到任务 -> 大脑(语言/理解)分析 -> 手(动作)直接执行。

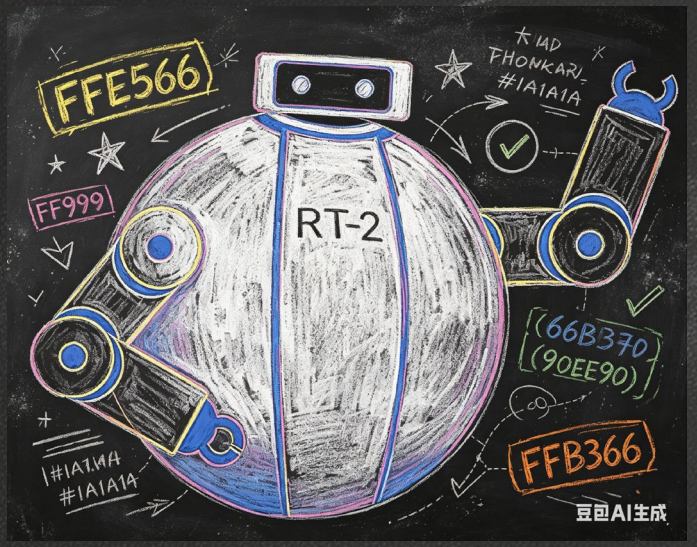
# VLA 代表工作

## Pi-0

Physical Intelligence · 通用机器人策略模型

打破传统机器人任务边界，将**视觉、语言、动作**三大核心模态进行统一建模。
具备极强的泛化能力，能适应从未见过的环境与任务指令。

核心亮点：
端到端的多模态融合决策，实现“感知-规划-执行”一体化。



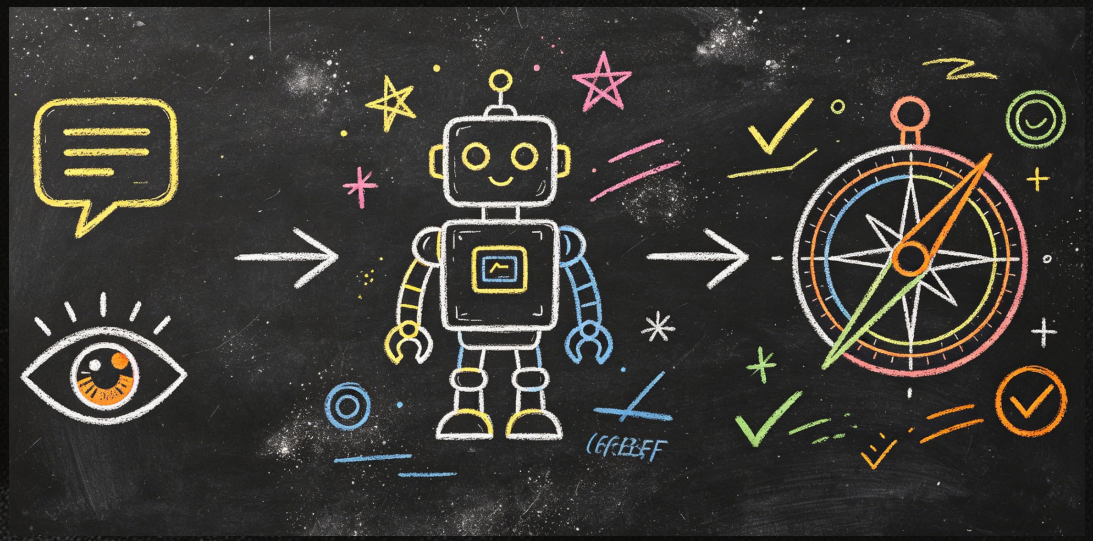

## RT-2

Robot Transformer (Google DeepMind)

突破性地利用**海量网络数据**（文本+图像）训练模型，将世界知识直接迁移并泛化为实体机器人可执行的物理操作技能。

# VLN — 视觉-语言-导航模型

## VISION-LANGUAGE-NAVIGATION MODEL



### 核心任务 Core Task

融合**视觉感知**与**语言指令**，实时规划出机器人从当前位置到达目的地的连续动作序列。

### 关键输入 Input Data

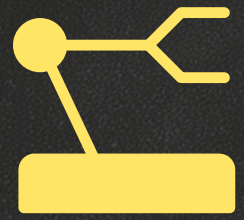
1.**视觉信息：**第一人称视角的连续图像或深度图
2.**语言指令：**自然语言描述的目标（如“去厨房拿水杯”）

### 决策输出 Output Action

离散的导航动作指令，例如：**前进、左转、右转、停止**等，指导智能体完成路径规划。

# VLA vs VLN — 操控 vs 导航

## VLA



### 核心关注 · FOCUS

Manipulation (双手精细操作) — 对物体的交互与控制

### 动作输出 · OUTPUT

精准的手臂/关节动作序列，执行复杂的肢体控制指令

## VLN

### 核心关注 · FOCUS

Locomotion (自主移动) — 在环境中的路径规划与探索

### 动作输出 · OUTPUT

连续的导航动作指令，实现从A点到B点的目标寻找

# 05

# 第五部分 / ACT 算法详解

ACTION CHUNKING WITH TRANSFORMER

CVAE架构 · Transformer编码器解码器 · 模仿学习

# ACT — 是什么?

## Action Chunking with Transformer

### 核心定义 / Definition

全称 Action Chunking with Transformer，是专为机器人操作任务设计的深度学习框架。

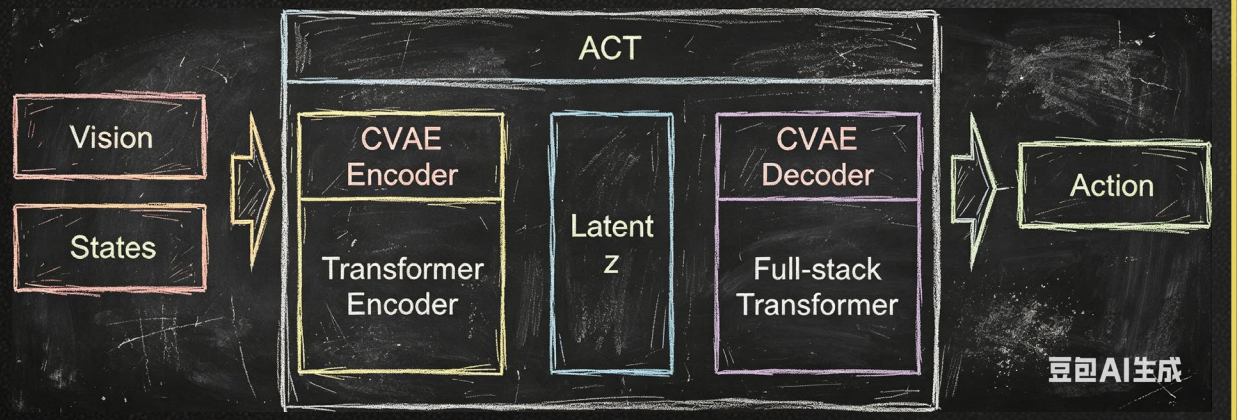
### 算法定位 / Positioning

一种端到端的具身策略算法 (Embodied Policy)，专注于解决复杂的 Manipulation (物体操作) 任务。

### 核心方法 / Methodology

基于模仿学习 (Imitation Learning)，通过观察并模仿人类专家的动作序列来驱动机器人执行任务。

输入：RGB视觉图像、机器人关节状态 |输出：连续的关节控制量、末端夹爪位姿




▶ *算法执行逻辑流程示意*

# ACT — 为什么需要它?

## ❌ 传统 Pick-n-Place 困境

传统机器人抓取依赖极其繁琐的流水线计算:
① CV 目标检测定位 → ② 三维点云计算姿态 → ③ IK 逆解规划路径。
这套流程对环境参数高度敏感，很难泛化到未知的新场景中。

流水线计算：检测 → 计算 → 规划

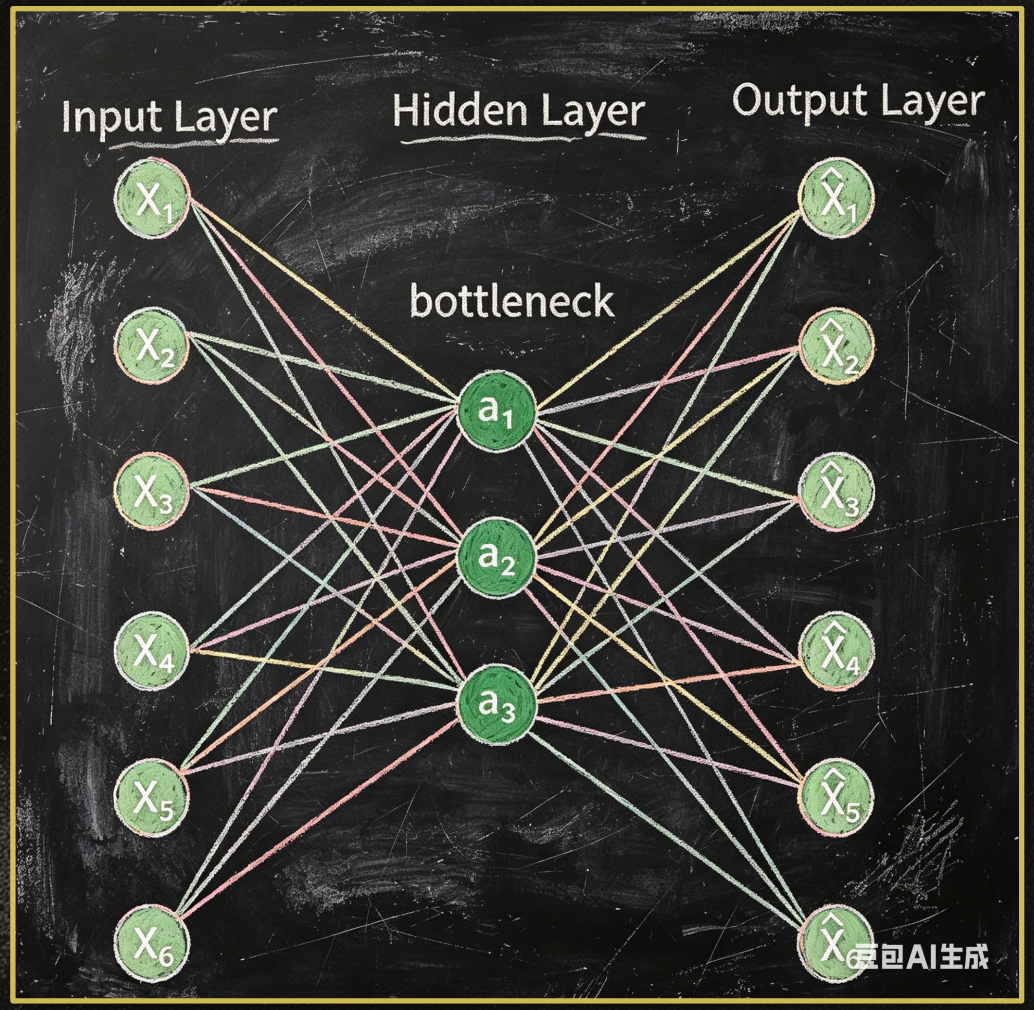
核心痛点:
过度依赖精确的环境建模与人工特征工程，缺乏对非结构化场景的灵活适应性。

## ✅ 人类 & ACT 核心思路

人类通过视觉直接映射到肢体动作，无需显式的中间坐标计算。ACT 算法正是模仿了这种高效的「身体智能」模式。

💡 ACT：从「感知」到「动作」的端到端直接映射

# 自编码器 (AE) 基础 / Autoencoder — 无监督维度约减




## 核心流程 / Process

数据输入 → Encoder (编码压缩) → Latent (隐层特征) → Decoder (解码重构) → 输出 ≈ 原始输入

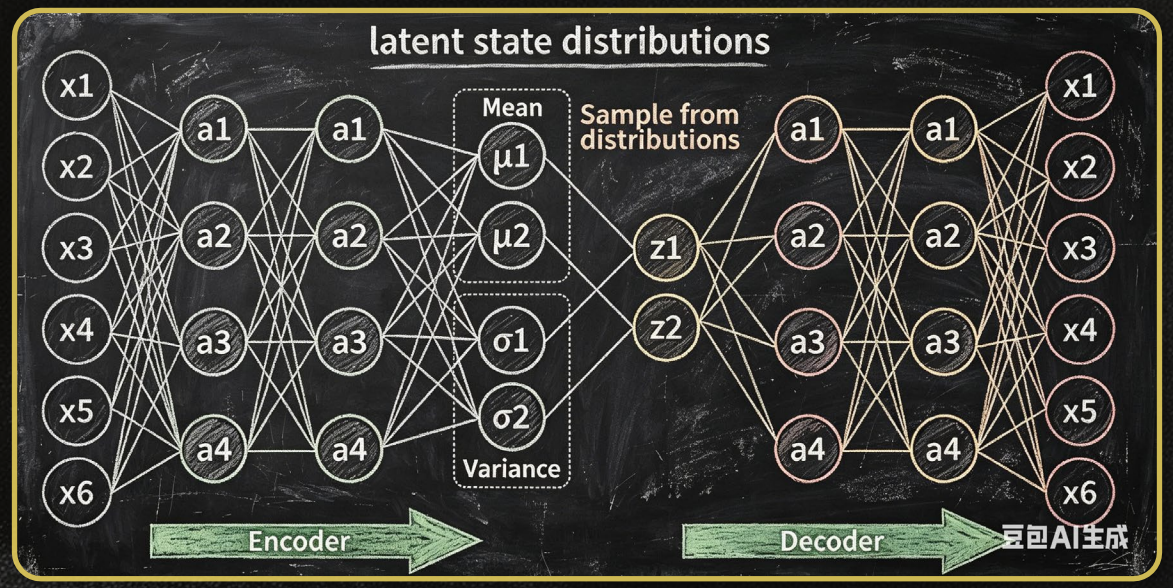
## 数学表达 / Formula

编码: Encoder(x) = z (获得隐层向量)
解码: Decoder(z) = $\hat{x}$ (重构输入)
目标: 最小化重构误差，即 $\hat{x} \approx x$

## 核心作用 / Function

本质是无监督的维度约减 (Dimensionality Reduction)，学习数据的低维子空间映射函数 f(x)。

# 变分自编码器（VAE） / 从固定 Latent 到分布建模




## 传统 AE 的核心痛点

Latent 空间是固定向量，生成过程缺乏随机性，无法灵活控制生成的多样性。

## VAE 的核心改进思路

将 Latent 建模为连续的多元高斯分布 (Gaussian Distribution)，引入概率生成能力。

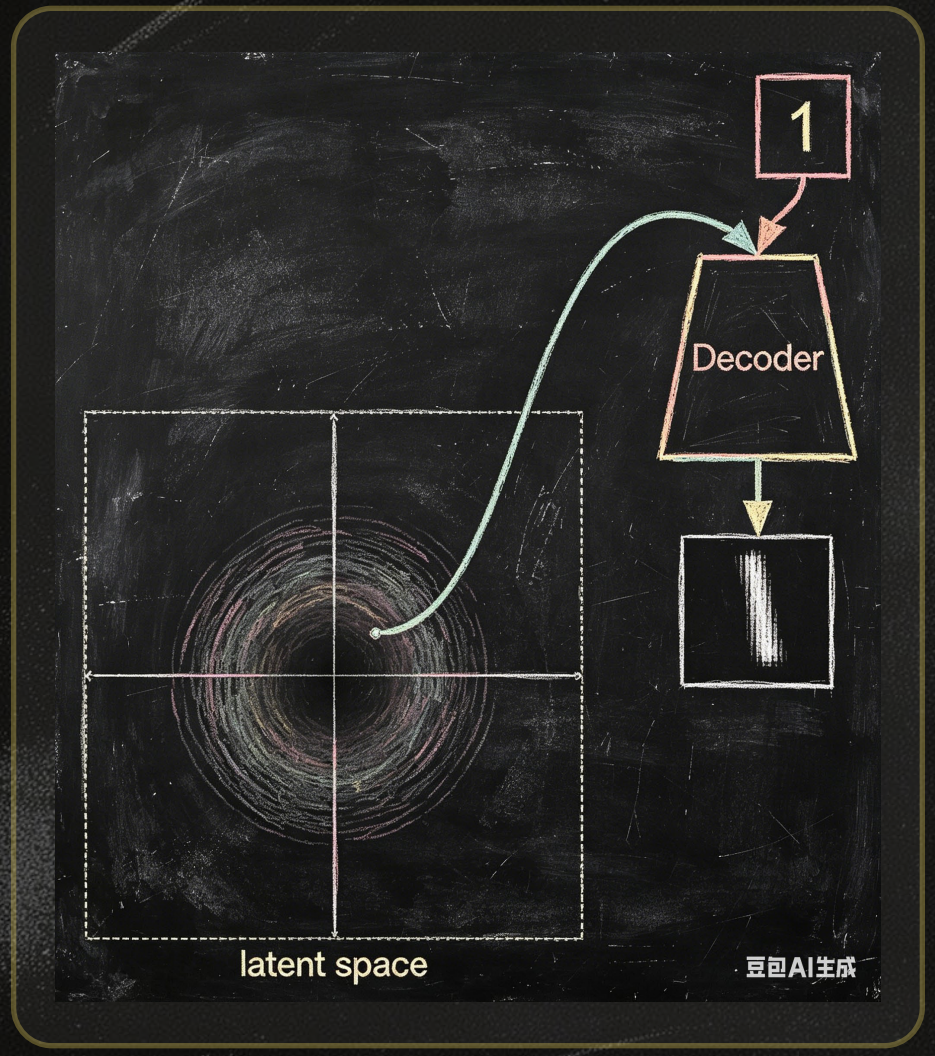
## 核心生成流程逻辑

Encoder 输出分布参数 μ(均值) 和 $\sigma^2$(方差)，基于 N(0,I) 采样得到 z，送入 Decoder 重构。

## Loss Function = 重构损失 + KL 散度

MSE (衡量重构误差) + KL (衡量与标准正态分布的距离)

# 条件变分自编码器（CVAE）/ 可控的生成 One-to-Many




## VAE 的局限性

标准VAE虽然能生成数据，但生成过程是随机的，**无法控制**具体生成什么内容。

## CVAE 的核心改进

在模型中引入额外的**Condition (条件)**输入，将生成过程从“不可控”转变为“可控”。

## 关键网络组件

包含**Encoder (编码器)**、**Decoder (解码器)**以及专门处理条件信息的**Condition Network**。

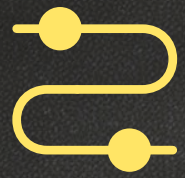
## One-to-Many 可控生成

通过输入不同的条件（标签、文本等），控制模型生成特定目标，实现一对多的生成效果。

# CVAE 在 ACT 中的作用

## 条件输入 (Input)

机器人当前的视觉图像（多视角）
与机械臂的实时关节状态



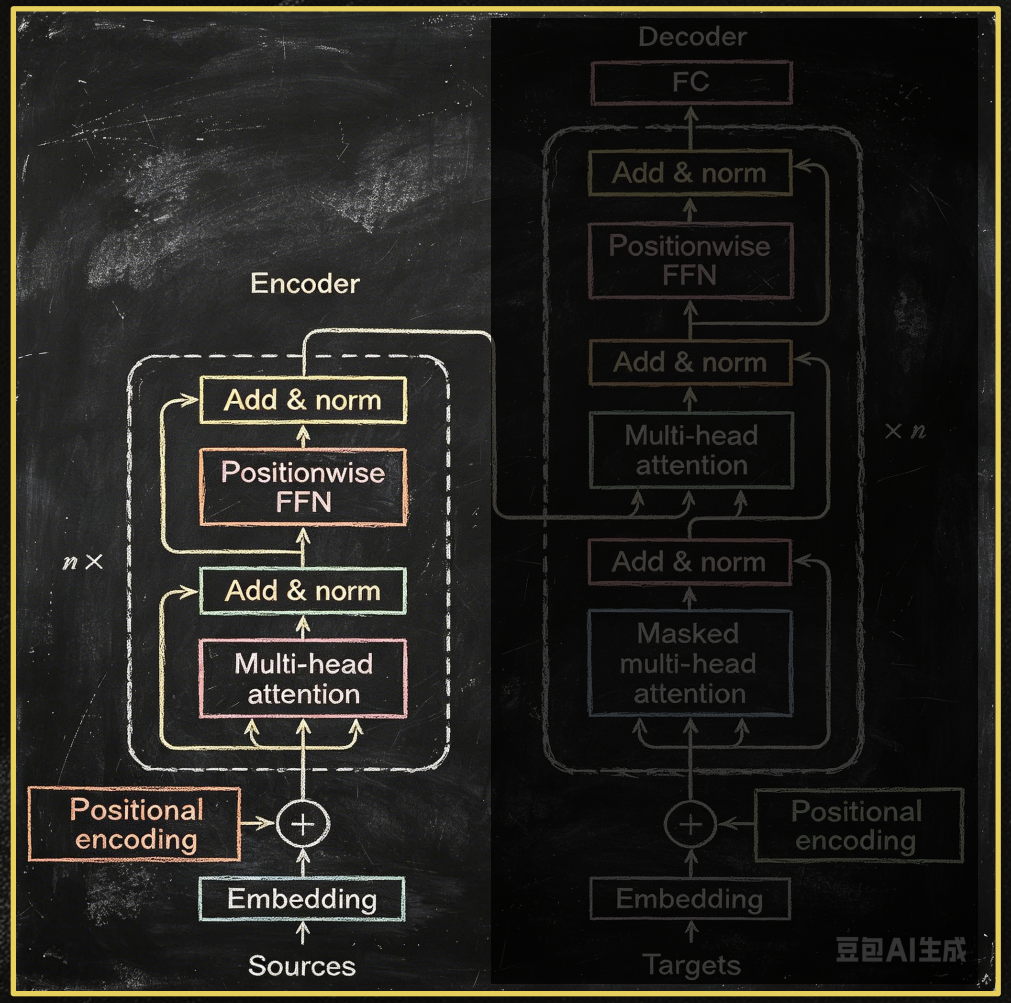
## 输出结果 (Output)

生成未来几个时间步的
动作序列 (Action Chunks)

## 核心机制 (Core)

在 Latent Space (潜在空间) 中
建立当前视觉状态与未来动作的映射
关系

# ACT 编码器 — Transformer Encoder




## 输入序列构造 (Input Sequence)

输入包含 k+2 个 Token：k 个目标动作序列、当前关节状态 Embedding，以及一个特殊的分类 Token [CLS]。

## 核心编码计算 (Core Calculation)

Transformer 处理后，提取 [CLS] Token 特征。通过一个线性层 (512→32) 映射，预测潜在向量 z 的均值 μ 和方差 $\sigma^2$。

## 潜在空间映射 (Latent Mapping)

生成的潜在向量 z 有效体现了从“当前视觉观测 + 状态信息”到“目标动作序列”的非线性动态映射关系。

# ACT 解码器 — Transformer Decoder

## INPUT / 输入条件

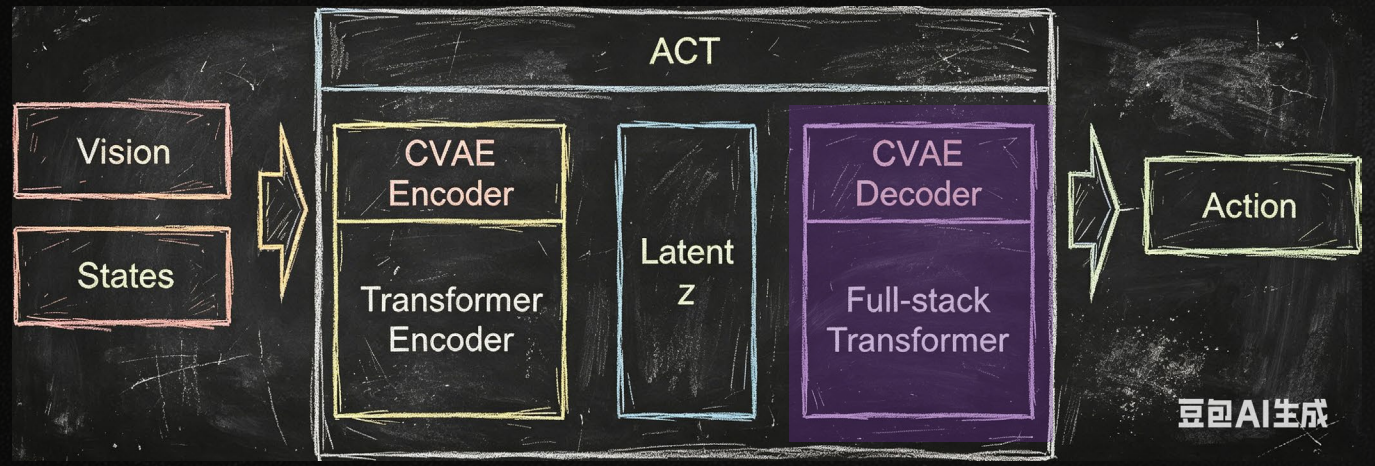
接收采样隐变量 z，结合当前观测图像与关节状态作为生成的初始 Condition。

## CORE / 核心解码映射

将隐变量 z 与观测信息在 Latent Space 建立映射，输出未来的动作序列 Chunks。

## ROLE / 时序建模能力

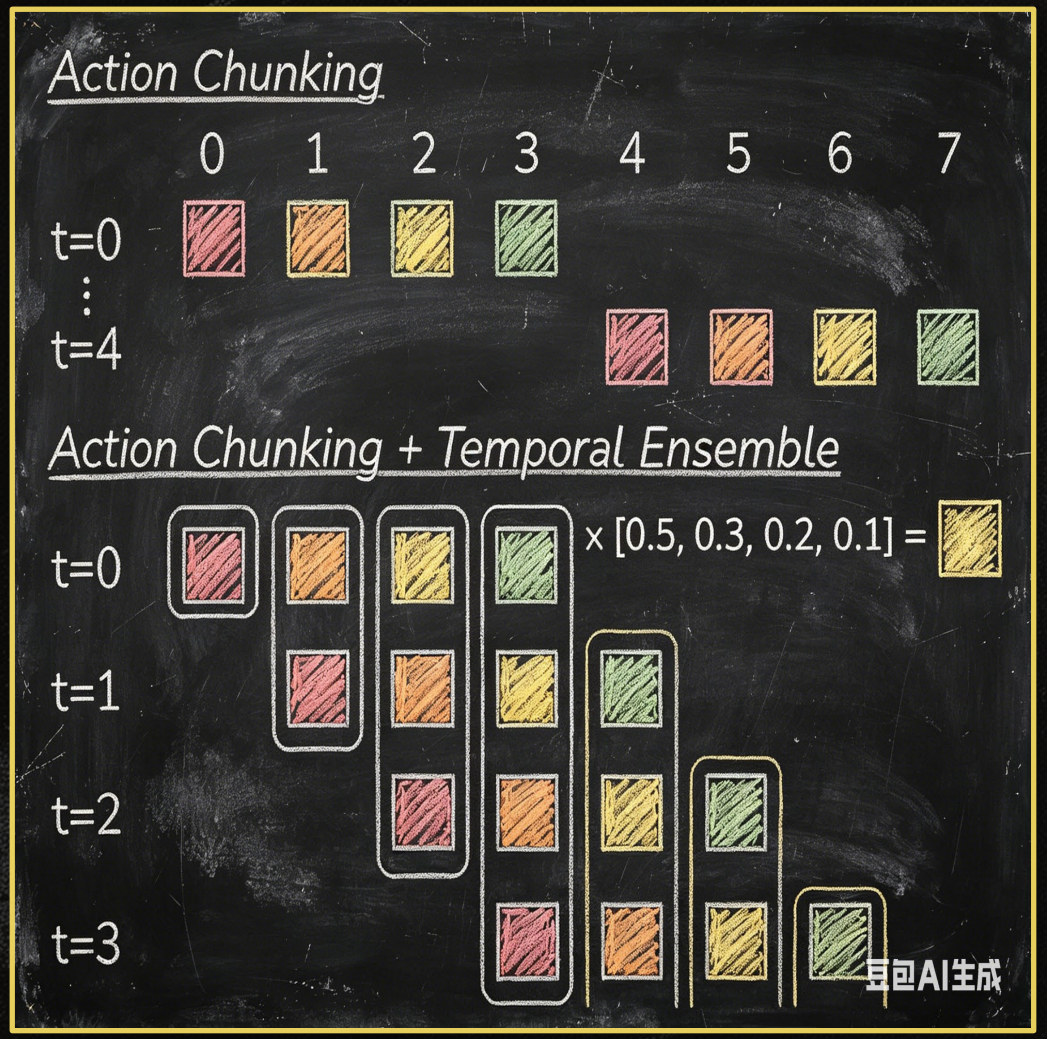
利用 Transformer 多头注意力机制，有效处理时序数据并捕捉长距离依赖。




## 💡 Attention Mechanism Visualization

图示展示了 Transformer 的核心 —— 多头注意力机制。在解码器中，它通过计算 Q, K, V 的相关性，让模型能够“关注”到输入序列中与当前预测最相关的部分，从而生成连贯、合理的未来动作。

# 动作分块 — Action Chunking




## 传统方法 | Traditional Step

每次仅输出单个动作 step，误差随时间逐步累积，导致整体时序连贯性差，表现不稳定。

## ACT 动作分块 | Chunking

每次直接输出动作序列 chunks，一次性预测未来多个时间步的动作，打破单步输出的局限性。

## 核心优势 | Key Advantages

有效减少误差累积，动作轨迹更平滑自然；更贴近人类连续行为模式，且能更好地适配 Transformer 的长时序建模特性。

# ACT vs VLA — 有何不同?

## VLA 范式

### 🎯 核心逻辑：视觉-语言指令驱动

模型运行依赖对自然语言指令的精准解析。必须将人类的任务目标转化为机器可理解的文本语义，串联起后续的视觉感知与动作生成逻辑。

#### 💬 典型任务指令

*“请识别桌面上的红色积木块，并将其精准地堆叠到蓝色积木块的正上方，同时保持整个结构的稳定性。”*

## ACT 范式

### 🎯 核心逻辑：场景状态（任务）驱动

🚫 无语言指令 · 仅接收【纯视觉 + 环境状态】输入

### 🦾 具身策略：模仿学习 (Imitation Learning)

采用模仿学习范式，直接通过观察人类演示的原始动作序列进行端到端学习。无需语义翻译，更贴近生物的本能反应与肌肉记忆。

# ACT 完整流程

01

## 编码器输入

接收 k 个目标动作、关节状态与 [CLS] 标志位，进行特征整合。

02

## 隐空间预测

利用 [CLS] 特征，通过神经网络预测潜在空间 z 的均值与方差。

03

## 采样隐变量 z

基于预测得到的均值与方差，从正态分布中采样得到隐变量 z。

04

## 解码器融合

解码器接收采样的隐变量 z 与当前环境观测，进行动作生成。

05

## 输出动作序列

生成并输出一段离散的动作序列 Chunks，作为机器人的执行指令。

06

## 闭环迭代执行

机器人执行动作，环境产生新的观测反馈，循环回到步骤1继续。

# ACT 的核心特点

## 动作 Chunk 化

输出连续的动作 chunks 而非单个离散动作，确保了动作执行的连贯性与流畅度。

## 多模态感知输入

模型同时接收视觉图像信息和机器人关节状态信息，实现多源数据的深度融合。

## 无语言指令依赖

不依赖人类语言指令进行控制，这是 ACT 模型与 VLA 模型最核心的区别所在。

## 端到端直接映射

建立从「视觉-状态」到「动作」的直接映射关系，大幅简化了网络的整体结构。

## CVAE 可控生成

引入 CVAE 变分自编码器，在生成动作的同时，实现了对动作多样性与可控性的调节。

## Transformer 时序建模

利用 Transformer 的自注意力机制，有效捕捉动作序列中复杂的长时序依赖关系。

# 06

# 第六部分 / 具身智能的挑战与未来

现存问题 · 发展趋势 · 课程总结

# 具身智能的挑战

## Sim-to-Real 迁移

仿真到真实物理世界的模型泛化能力，是落地应用的核心难点。

## 长程任务规划

建立“感知-决策-行动-反馈”的完整闭环，保障复杂任务的稳定运行。

## 数据效率

应对真实机器人数据稀缺问题，在有限数据下实现高效的策略训练。

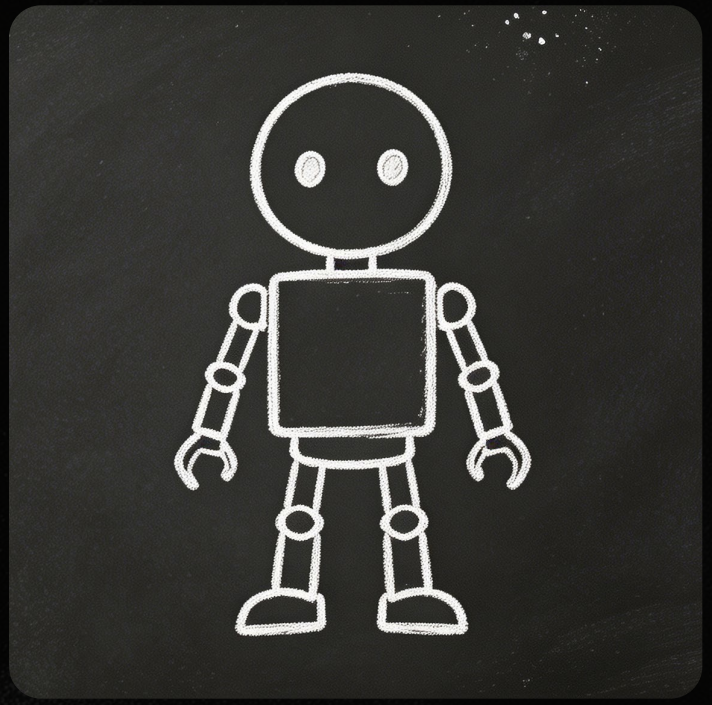
## 安全可控

确保物理世界中，机器人能与环境及人类进行安全、可靠的动态交互。

## 多模态融合

对视觉、触觉、力觉等多源异构信息进行统一建模，实现对环境的深度理解。

# 未来发展方向



EMBODIED AI

## 多模态大模型进化

多模态大模型机器人(VLA)持续进化，深度融合视觉、语言与行动指令。

## 通用具身智能体

打破任务边界，致力于实现“一种模型，解决多种复杂任务”的通用能力。

## 高质量仿真平台

构建高保真虚拟物理环境，大幅降低试错成本，加速智能体的训练迭代。

## 硬件技术革新

低成本、高自由度的人形机器人硬件普及，为具身智能落地提供物理基础。

## 持续学习与成长

赋予机器人长期记忆能力，使其能在环境交互中像人类一样持续学习进化。

# 课程总结

具身智能≠离身智能替代，而是对其功能的重要补充与拓展。

VLA、World Model 等前沿技术，目前只是实现具身智能的众多有效路径之一。

智能并非凭空产生，而是源于智能体与物理世界持续交互的过程中。

核心算法公式：
ACT = CVAE + Transformer + 模仿人类动作序列

优秀的具身策略（Policy）中，往往隐含地覆盖了一部分具身规划能力。

核心结论：
具身智能是通向 AGI (通用人工智能) 的必要且关键组成部分。

# 谢谢！

Questions & Discussion / 问答环节