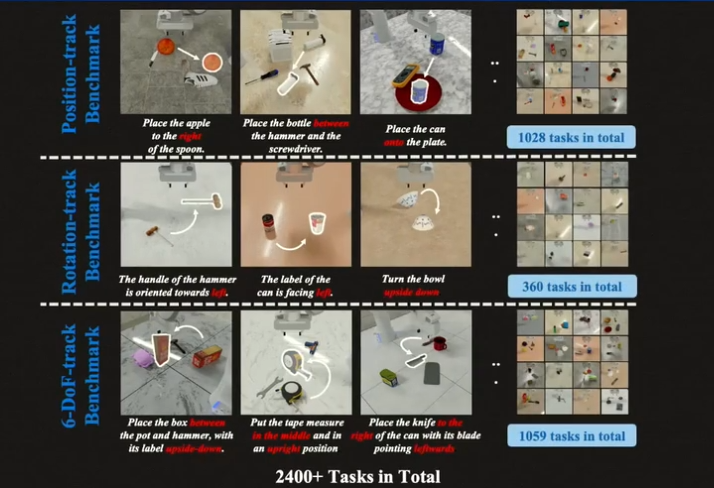
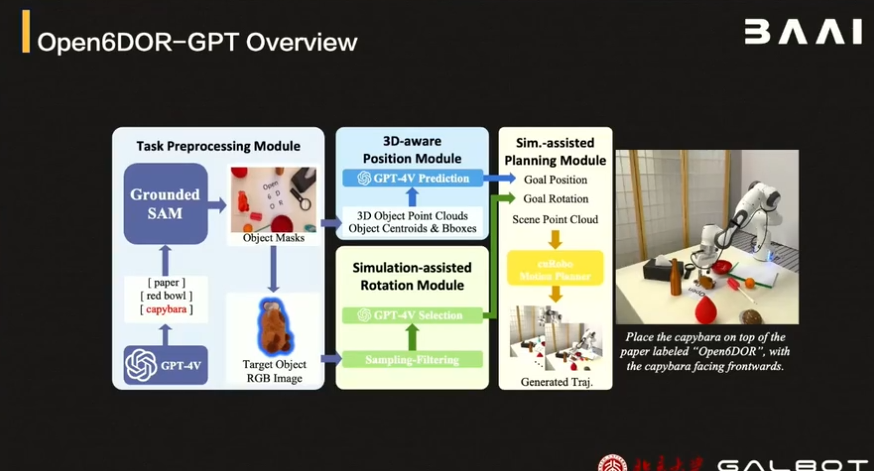


可以把每一盒药的纹理建的非常清楚，解析出药品内容，最终实现一个三维语义地图，然后走到这个药盒子附近拍照然后再进行药品选择抓取（必要性）：和机器人动作匹配，抓取

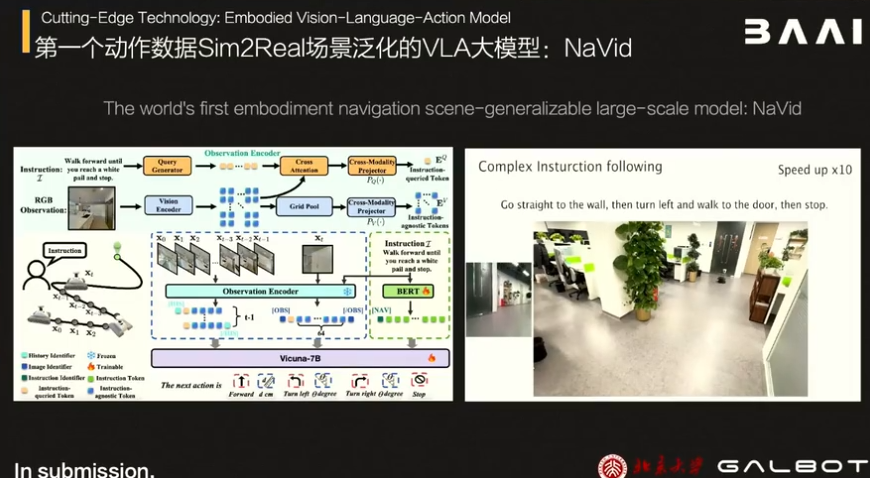
除了抓，能不能放：

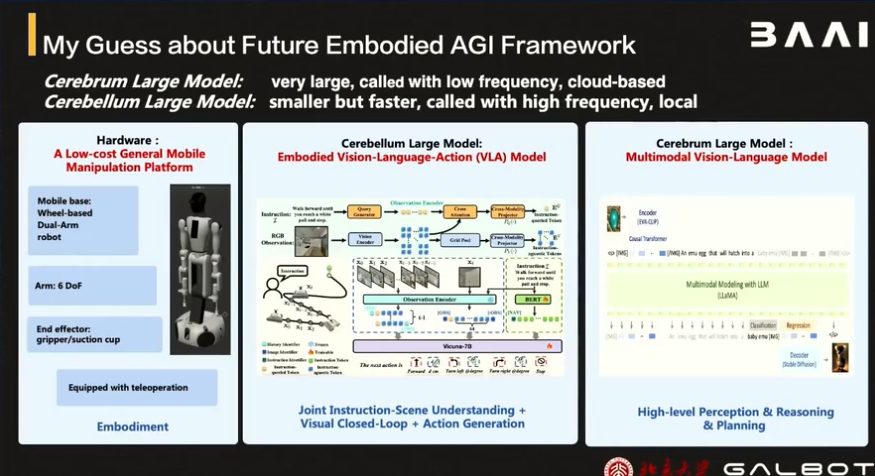


谷歌：把。。放到xx位置，既可以说朝向，又可以说位置，还可以说比如：朝上、朝下



首先SAM，然后GPT4V决定goal position,同时，进行朝向判断决策选择物体要朝向的位置的图像并生成goal rotation，之后就是能够完成：把卡皮巴拉放到写有open5DOR的纸上面，并把卡皮巴拉朝前



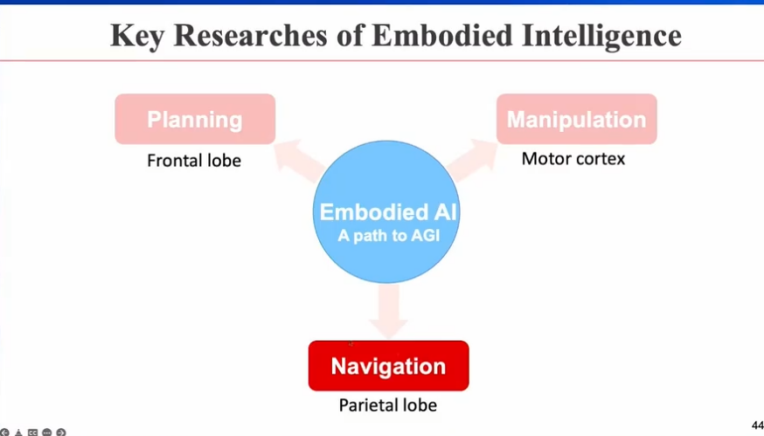


关键：仿真数据的重构对实现具身智能0-1的过程很重要

利用大模型专家辅助完成任务。

1. **具身智能关键技术研究：操纵、决策、导航---董豪**

谷歌RT1 有缺陷：放到另一个房间里就失效了】



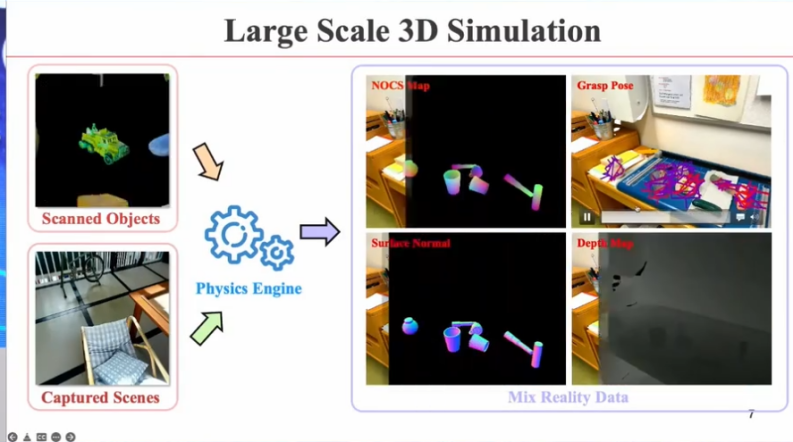


**具身智能大大受限在数据本身**

我们一上来就要做端到端吗？其实不是，去看人体大脑的结构，是有不同脑区，不同脑区负责不同的功能，比如有视觉、听觉、嗅觉感知，空间导航、决策规划，手、腿运动相关

然后以上这些其实可以发现在大模型当中已经可以做到，比如说感知能力，大模型通过互联网就可以知道。但是关于手部、腿部能力，受限互联网没有数据。

**机械臂抓取：**



3d物体，扫描环境，有了物体、环境，之后在物理引擎当中得到AR，直接计算位姿，利用仿真，可以无穷尽的生成数据

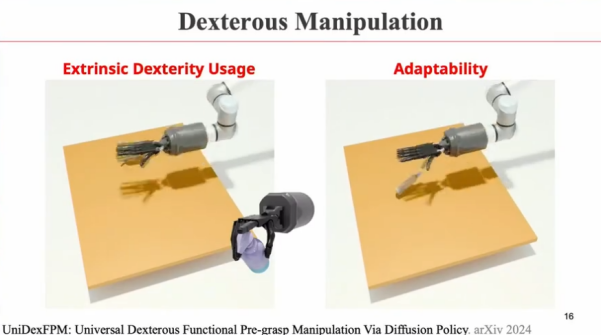
姿态估计：为了更好的放置

**灵巧手：我们思考人是怎么抓取的：首先调整物体姿态方便手去抓取的**



有的时候三指就可以抓取了，那什么时候需要5指呢？

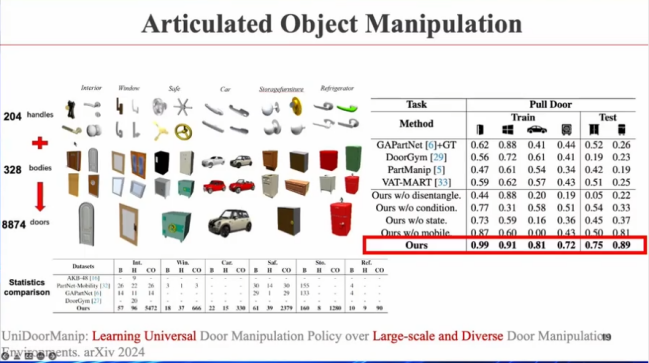
为后续任务做铺垫：



用**强化学习**



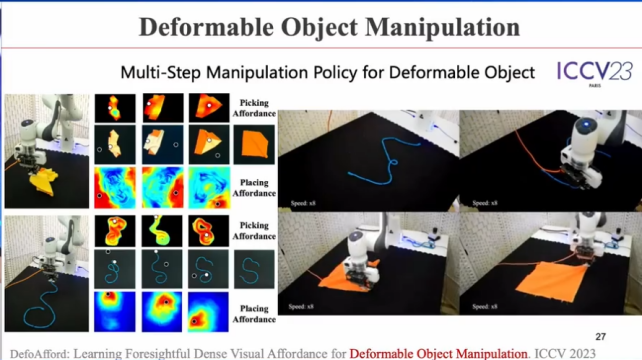
仿真的边界在哪里：做了一个实验，仿真不同的门把手和门版面，然后组合出不同的门，发现在千级左右的门就可以达到很好的效果

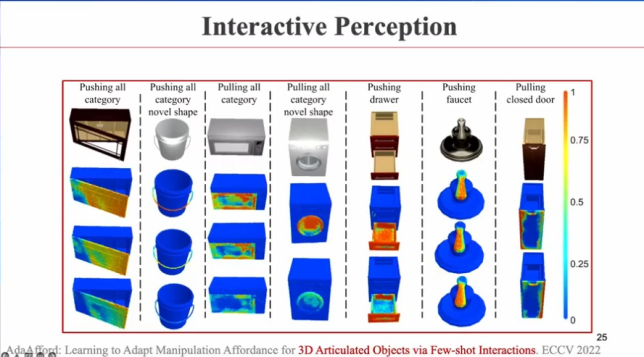


即将开源:sim2sim能更好的完成规模化测试



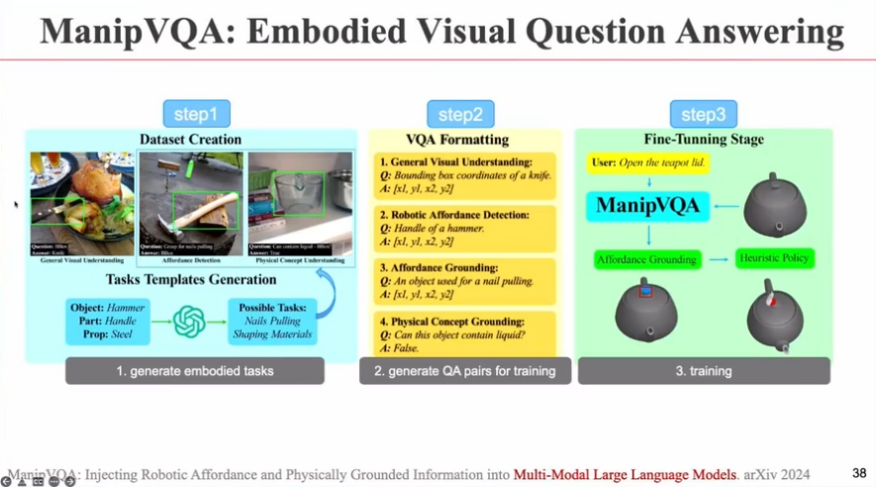
那么目前仿真比较好的就是视觉，可以做大量数据增强







确定位姿，之后相机随便摆，实时确定位置，姿态估计，进行视觉闭环控制

**VQA**  

利用大模型生成可能会有的任务然后生成数据

任务规划有大模型，所以不是难点

具身智能的难点还是在机械臂的操作上面

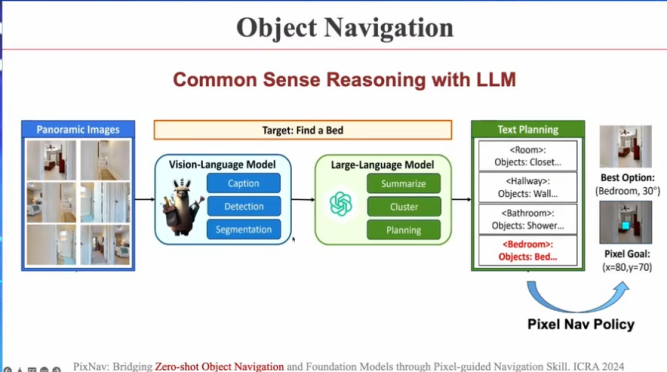
如何让机械臂知道什么是整洁：

输入很多图像，让其明白“整洁”是怎样的分布。

**具身导航**

**物体导航ObjectNav**

在环境中探索然后给你



利用大模型做这样的事情，多模态视觉语言大模型完成感知理解的任务，然后用大语言模型完成规划任务，并且预测下一步走到当前画面哪个点-》实现物体的导航

**视觉语言导航VLN**

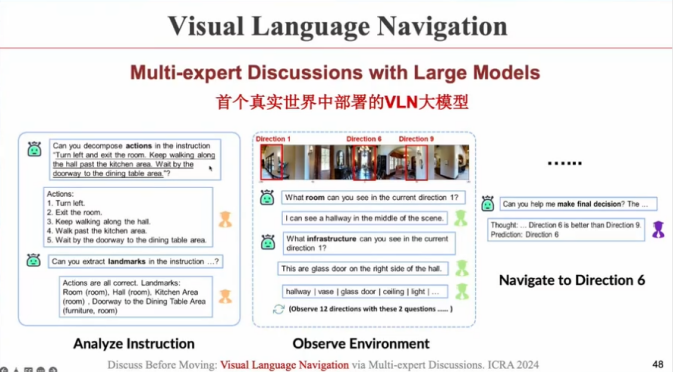
跟随人类的语言进行行程改变



第一次在真实世界部署的VLN大模型，过去都是在仿真房间里学，往往真实世界不行，往往换一个型号的相机，相机高低都会影响效果

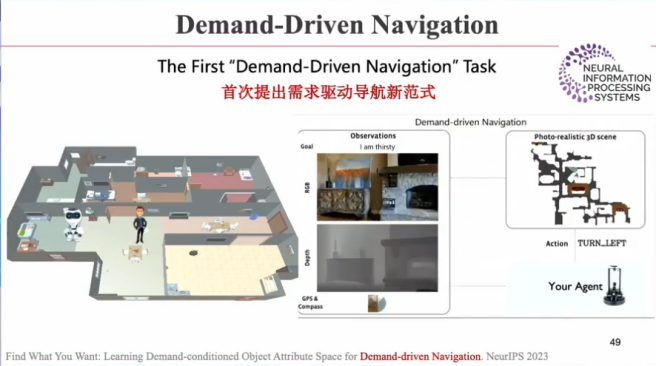
我们利用大模型自问自答的方式：一个专家把人类命令拆解，另一个转变成要行动的地方和方位，最后一个专家决定往哪里走。

并不需要训练模型，只需要类似内心对话一样得到结论



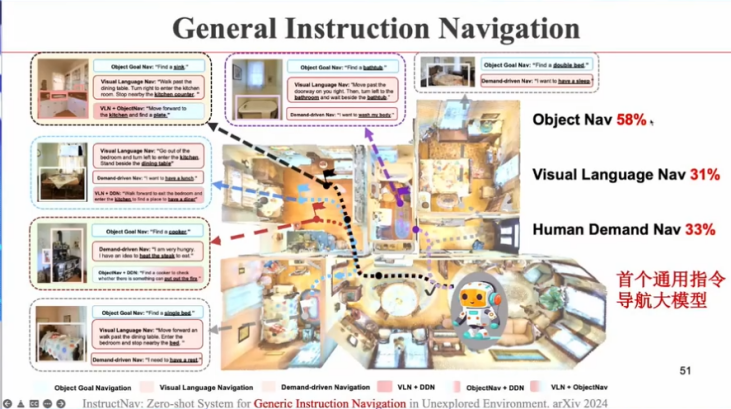
**需求驱动导航新范式DDN**

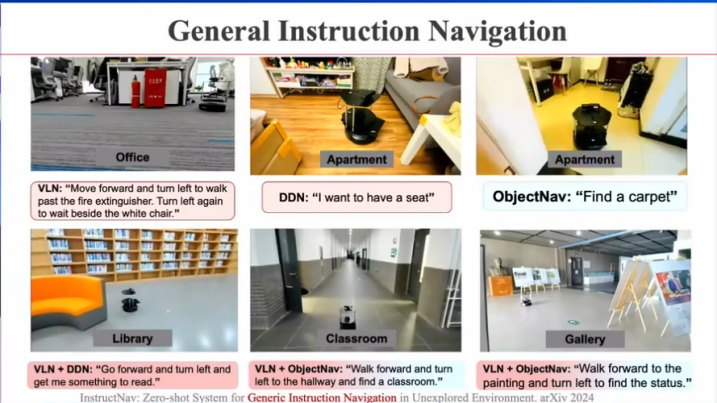
以往都是提出一个要求，比如我要水，还得告诉机器人往哪走那走。不如直接告诉他我们渴了，然后机器人去找水，如果没有找到说不定还会给我拿一个西瓜这样子过来



**通用指令导航大模型**

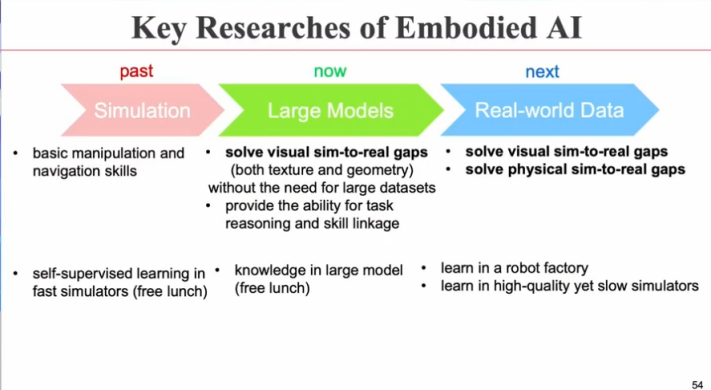
将之前多种导航范式全部包含





一个模型可以做到所有任务，看这里的任务完成度都很低是因为机器人没有完成建图的任务，第一次来你家然后完成任务，如果完成建图那么成功率在90%

我们为什么现在会使用大模型而不是过去的仿真了：是因为大模型可以很好的解决sim to real 的问题，而且还可以提供任务编排的能力，那么下一步的研究趋势就是：real word data,因为物理层面的sim to real还是很难



导航：目前会用gpt-4v,阿里的通义千问

过去的仿真数据还有意义吗？在很多的技能方面是很有必要的，短期内还在做，所以说仿真技术会越来越好，减少真实世界成本--追寻数据金字塔

真实数据采集，工作量成本会不会特别高：有成本，很难采集到多样化的数据