洞见： https://www.4paradigm.ai/

解读 NVIDIA 公司的 Jim Fan 博士对 2025 年机器人技术现状的“年终总结”：Jim Fan of Nvidia posted three lessons for robotics in 2025

这段推特博文是**樊锦博士**作为一线智能机器人研究者，在英伟达实验室里被硬件折磨、被数据误导、被算法瓶颈困扰后的真情流露。以下是对这三大教训的深度解析：

---

1. Hardware Ahead of Software / 硬件技术领先于软件技术，但硬件的可靠性严重限制了软件的迭代速度。

Jim Fan 博士指出了一个极其尴尬的现状：

硬件功能卓尔不凡： Optimus、Atlas、Figure 等机器人的自由度（DoF）和动力学性能已经非常强悍，甚至超出了 AI 模型的应用极限。换句话说，机器人的身体性能远超其 AI 大脑的控制能力。

硬件可靠性乏善可陈： 软件迭代依赖于大量的测试，但硬件的过热、电机故障、各种奇怪的固件问题每天都困扰着我们。由于没有“自我修复”能力，一旦电机烧毁或固件崩溃，整个研发进度就会停滞。这导致机器人研究依然是极度依靠运营团队的维护。

Jim Fan 博士其实提到二层含义：首先提到的*机器人硬件寿命和可靠性。特斯拉 Optimus 机器人，因其关节过热也引起广泛的关注，这也是****人形机器人面临的普遍技术难题****，主要源于电机、减速器高密度集成在狭小空间内产生的****高热量难以散出****，导致扭矩下降、精度降低甚至触发保护停机，另外据 2025 年 10 月的报道，Optimus 机器人灵巧手在快递分拣工作中的寿命仅为 6 周。单只手的成本超过 6000 美元，加上其他易损部件，维护成本很高。由此可见，提高机器人的“健康”和长期可靠性是实现大规模商业应用的关键。第二，由于硬件的有限寿命和不可靠导致 AI 软件无法快速迭代。软件和硬件是“鸡和蛋的关系”，短期是软件决定硬件，长期是硬件决定软件。结合 Jim Fan 博士在 2025 年 5 月份的演讲，我们可以推测 NVIDIA“将”（或者已经正在做）通过仿真驱动实现机器人在物理世界中的零样本迁移能力。*

*简而言之，NVIDIA 将****通过在虚拟环境中模拟物理世界，能够以更低的成本、更高的效率生成大量数据，从而训练出更智能的机器人软件。至于硬件问题嘛！这个不是 NVIDIA Jim Fan 博士能够搞定的。***

---

2. 机器人领域的基准测试之乱

洞见： https://www.4paradigm.ai/

Jim Fan 博士抨击了当前机器人学界缺乏广泛接受的基准评测的方案。他呼吁 2026 年必须建立像科学实验一样的严谨评估体系。

- **缺乏“统一基准测试方法”：** 大语言模型有 MMLU 这种公认的跑分，但机器人领域每个人都在自己定义的任务、自己搭建的场景里拿第一。
- **演示骗局（Cherry-picking）：** 社交媒体上看到的近乎完美的机器人表演视频，背后可能是 99 次失败后的唯一一次成功。这种缺乏**可重复性**的现状，掩盖了行业真正的技术瓶颈。

Jim Fan 博士其实提到现实机器人行业的二层问题。 首先，缺乏“统一基准测试方法”其实是因为机器人是新鲜事物，很多机构都投入资源来研发。如果有一套标准的基准测试方案，那么大家就可以横向和纵向进行比较。但是机器人的研发日新月异，如何可以达成共识呢？即使有“共识”，你不担心有人“投机取巧式作弊“吗？另外，统一基础测试某种情况也会变成唯一的导向，最终造成”顾此失彼“。例如，高考数学分数高，就一定保证职业能够成功吗？第二，演示骗局这个就是见仁见智了。如果只有百分百成功才是成功，那些百分之一的成功难道就是失败吗？机器人行业现状是从 0 到 1，也就是从无到有，局部成功，阶段性的成功也是成功呀！

---

3. VLA 路线的“底层逻辑缺陷”

Jim Fan 博士认为基于大语言模型（VLM）改装的机器人大脑（VLA）可能有很大挑战：

- **预训练目标偏差：** VLM 的目标是“聊天”和“看图说话”，它的参数里装满了历史知识和语言逻辑，而不是物理世界的重力、惯性或摩擦力。
- **视觉细节的丢失：** 传统的视觉编码器（Visual Encoder）为了提高效率，会把一张图片抽象成“那是一把剪刀”，但机器人需要知道的是“剪刀尖端在那 1 毫米的缝隙里”。VLM 在抽象化的过程中，把精细物理操作所需的低层细节（Low-level details）全丢了。
- **他的解决方案：视频世界模型（Video World Model）。** 与其教 AI 读书说话，不如让它通过海量视频学习物理规律——比如球掉下去会弹起，玻璃撞击会碎。这种**对物理世界的预测能力**，才是机器人政策（Policy）的真正基石。

Jim Fan 博士其实提到现实机器人行业的发展轨道，他认为机器人应该侧重通过机器视觉来理解物理运动和协调，而忽略掉与基本运动无关的知识。或许，机器人应该把自己当成

洞见：https://www.4paradigm.ai/

“人类的婴儿“，只关注最基础的需求也就是物理运动，忽略掉那些高级视觉带来”人文知识“。短期看这个观点是对的，因为目前还在如果解决物理运动的阶段。但是长远看这个观点无疑是错误的。我们大胆假设，如果机器人只需物理运动，长此以往，这就是带来最最核心的问题，缺乏”人文知识“的机器人永远是”器“ 和”物“；人类最需要的乐于助人，体贴入微，力大无穷的”机器人“。

原文阅读：

https://x.com/DrJimFan/status/2005340845055340558

延展阅读 1

Jim Fan 本科毕业哥伦比亚大学计算机系，后在斯坦福大学视觉实验室获得了博士学位，师从李飞飞教授。他的研究领域十分广泛，包括了多模态基础模型、强化学习以及计算机视觉，曾实习于谷歌云 AI、OpenAI、百度硅谷人工智能实验室等知名组织。

Jim Fan 目前在英伟达公司领导 AI 相关研究，其团队正致力于开发“Project Groot”，这是公司为创建人形机器人基础模型所做的努力。

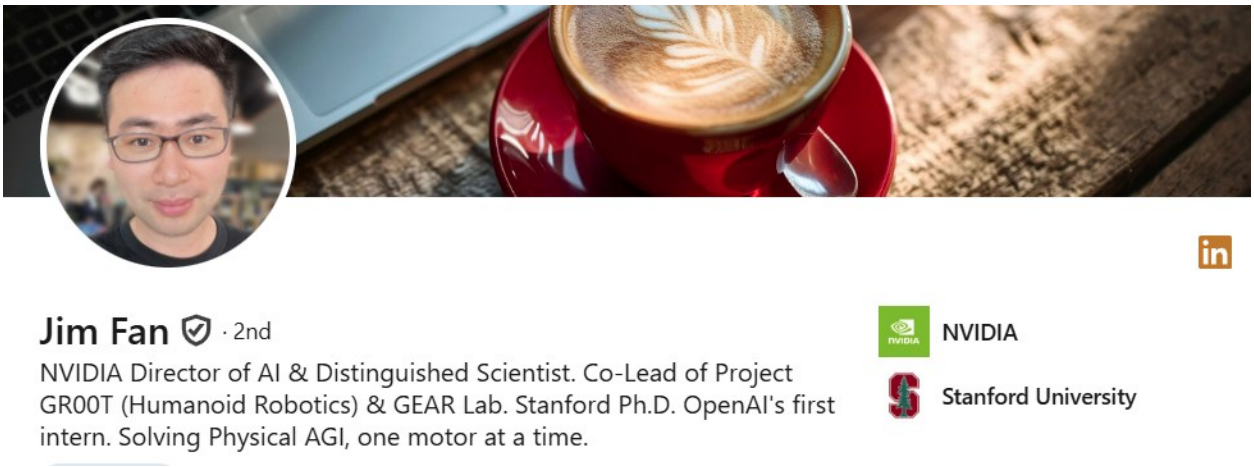

Fig.1. Jim Fan LinkedIn

延展阅读 2

https://developer.nvidia.cn/blog/accelerate-generalist-humanoid-robot-development-with-nvidia-isaac-gr00t-n1/

NVIDIA Isaac GR00T 通过提供开源的 SimReady 数据、仿真框架（如 NVIDIA Isaac Sim 和 Isaac Lab）、合成数据蓝图和预训练基础模型，能够帮助解决这些挑战并加速通用人形机器人的开发。

NVIDIA Isaac GR00T N1 的特点和优势

NVIDIA Isaac GR00T N1 是世界上首个用于通用人形机器人推理和技能的开源基础模型。这个跨实体模型接受包括语言和图像在内的多模态输入，以便在各种不同的环境中执行操作任务。

GR00T N1 基于一个庞大的人形机器人数据集进行训练，训练数据还补充了通过 NVIDIA Isaac GR00T Blueprint 生成的合成数据，以及来自互联网的大量视频数据。它可以通过后训练适应特定的实体、任务和环境。现在，开发者可以通过 Hugging Face 上的开源 NVIDIA 物理 AI 数据集免费获得其中的部分数据。

GR00T N1 模型架构

受人类认知原理的启发，GR00T N1 基础模型采用双系统架构：

- 视觉-语言模型（系统 2）：这个系统基于 NVIDIA-Eagle 和 SmolLM-1.7B，是一个方法论思考系统。它通过视觉和语言指令解释环境，使机器人能够对其环境和指令进行推理，并规划正确的行动。
- 扩散 Transformer（系统 1）：这个动作模型生成连续动作以控制机器人的运动，将系统 2 制定的动作计划转化为精确、连续的机器人运动。

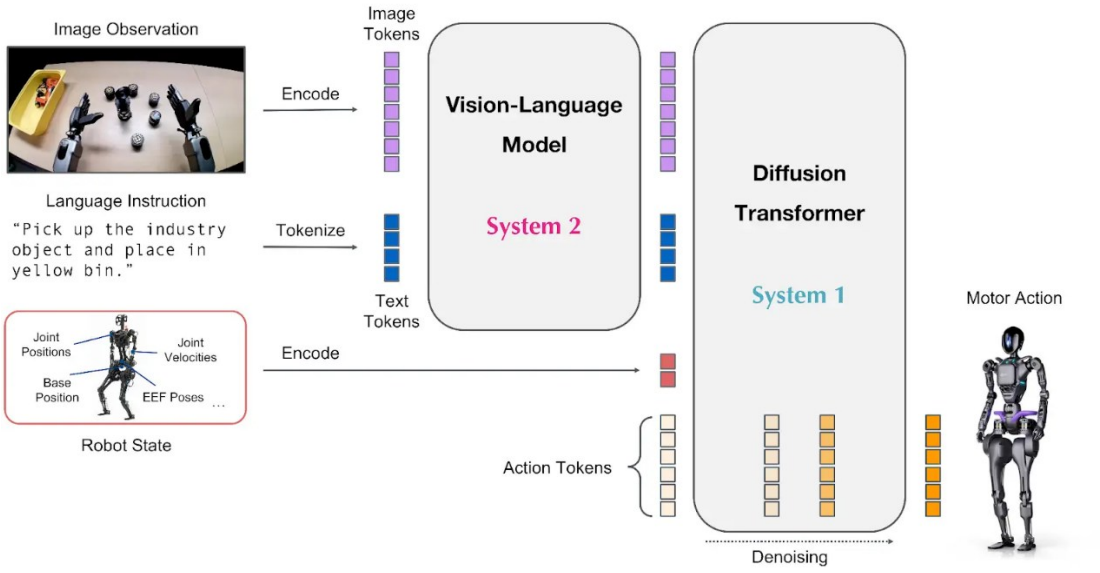

Fig.2 https://www.analyticsvidhya.com/blog/2025/03/nvidia-isaac-gr00t-n1/

洞见： https://www.4paradigm.ai/