从世界模型到世界智能体

From World Models to World Agents

斯坦福 CS292 / EE292R《AI and Data for Science》2026 秋季讲座

讲者：Gordon Wetzstein，斯坦福物理与空间 AI 实验室负责人

时间：2026 年 9 月 30 日 · 原课时长 54 分钟



原课开场演示：Google DeepMind 的 Genie 3。画面不是游戏引擎渲染的，是模型跟着方向键一帧帧现场生成的。

整理：心晴九天

本笔记为个人学习整理，内容以原课为准

OVERVIEW

一页看懂这堂课

大语言模型学的是怎么说话，世界模型学的是世界下一秒会变成什么样。这堂课从原理讲到难题，最后落到怎么用它训练能在工厂里干活的机器人。

它是什么 → 怎么造出来 → 三种用法 → 三个难题 → 变成 Agent → 未解问题

六句话记住主线

1. **它是什么**：一个实时的、跟着动作走、一帧接一帧往下画的视频生成器。比传统模拟器更像真的，但没有可以编辑的 3D 状态。
2. **怎么造出来**：扩散模型，从噪点一步步去噪出画面；网络用的是 Transformer；每一帧都参考前面几帧，以及动作、文字等条件。
3. **三种用法**：开环（游戏，自由但会越滚越歪）、闭环（机器人，有真摄像头拉回现实）、锚定开环（自己画，但靠地图这类不变的锚稳住）。
4. **三个难题**：记不住（上下文太短）、多人共处一个世界、越画越歪（曝光偏差）。各有对应的解法。
5. **变成 Agent**：带动作标注的视频只有约 2000 小时，普通视频约 20 亿小时。解法是先生成"成功完成任务"的视频，再用逆动力学反推动作。
6. **未解问题**：更长的上下文、和视觉语言模型怎么分工、去噪步数、检验员、世界动作模型怎么设计。

> **一个整理者的观察**。世界模型碰到的坎，和大语言模型几乎一一对应：上下文不够长、自己的错误越滚越大、需要被真实世界拉回来。大模型这几年迈过这些坎的办法，很可能就是世界模型接下来要走的路。

讲者本人的两句提醒

"生成模型一定会幻觉，这本来就是它被训练去做的事。"关键是让它一直贴着真实的观察。 53:00

现有的世界模型还不够好，真要卖给客户的机器人，他不会用它们来训练。 51:00

GLOSSARY

概念词典

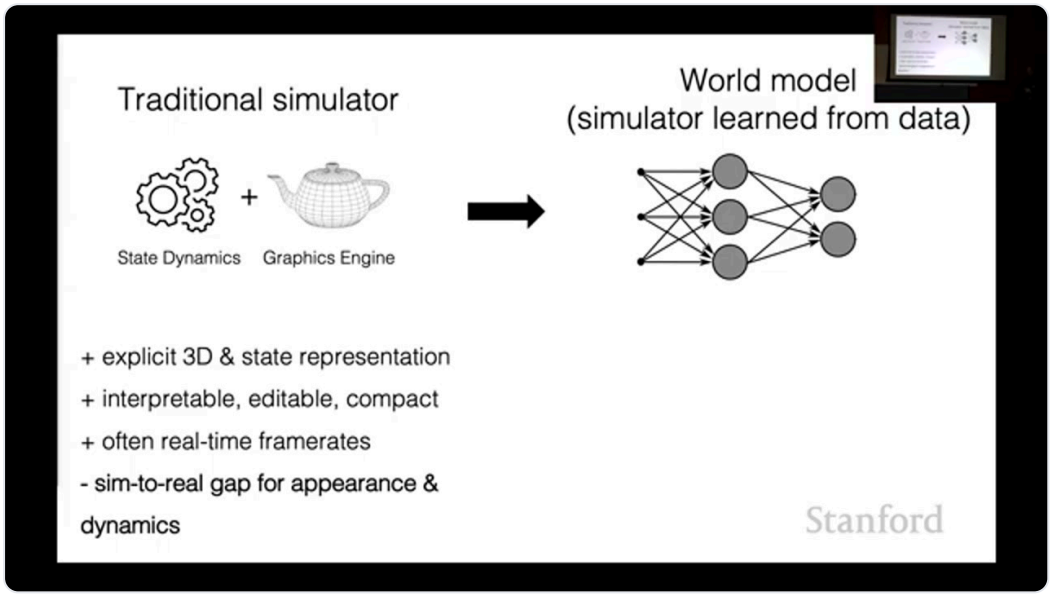
按出现顺序排列。右栏是一句话的白话解释，括号里是原课时间点。

概念	白话解释
世界模型 World Model	从视频里学会"世界下一秒会怎样"的模型。讲者的定义：实时的、动作条件的、自回归的视频生成器。（16:50）
仿真和现实的差距 Sim-to-Real Gap	传统模拟器里的画面和物理不够真，在里面练好的东西到真实世界常常不灵。（04:30）
扩散模型 Diffusion / Flow Matching	从纯噪声出发，一步步去噪生成画面；训练时反过来，给真画面加噪，让模型学习"往哪个方向去噪"。（07:00 – 11:30）
潜空间 Latent Space / VAE	先用变分自编码器把高清画面压缩成小得多的表示，在这上面去噪，最后再还原，省算力。（11:50）
扩散 Transformer Diffusion Transformer	负责去噪的网络，和大语言模型同一类结构，能靠堆数据和算力扩展。（12:30）
上下文帧与 KV 缓存 Context Frames / KV Cache	生成下一帧时参考前面几帧，用的是和大模型记住前文一样的缓存机制。（14:00）
世界动作模型 World Action Model, WAM	不仅预测未来画面，也预测 Agent 接下来该采取的动作。（16:30）
开环 / 闭环 / 锚定开环 Open / Closed / Anchored Open Loop	参考的是自己画的帧 / 真实摄像头的帧 / 自己画的帧加一个不变的锚（如地图）。（17:00 – 19:30）
曝光偏差 Exposure Bias	训练时看真数据，推理时看自己生成的数据，误差逐帧累积。大语言模型也有这个问题。（33:10 – 34:20）
FramePack 上下文打包	越久远的帧，存得越低分辨率、越稀疏，打包成一张图交给模型，用有限算力记更久。（24:00）
空间记忆 Spatial Memory	每帧估计深度，融合成全局 3D 点云地图，回到老地方时按地图把结构画对。（25:30）
DAgger / BAgger Dataset / Backwards Aggregation	DAgger：人类把跑偏的车拉回来，纠正过程变成训练数据。BAgger：把模型自己画坏的视频倒放，得到"从坏变好"的示范。（34:20 – 35:50）
策略 Policy	机器人的大脑：看到什么，下一步做什么，形成感知—动作循环。（37:00）
模仿学习 Imitation / Supervised Learning	用带动作标注的视频直接教机器人，是现在的主流，瓶颈是数据太少。（38:40）
逆动力学 Inverse Dynamics	已知前后两个画面，反推中间做了什么动作。数据效率很高。（41:00 – 42:30）
缩放定律 Scaling Laws	模型和数据足够大，才会出现思维链、上下文学习这类涌现能力。讲者认为机器人也要走这条路。（40:30）
少步蒸馏 Few-step Distillation	把原本要 30 – 50 步的去噪，压缩到一两步，以实现实时生成。（45:30）

PART 1

一 它是什么

传统的模拟器，也就是游戏和电影特效用的那一套，由两部分组成：一个状态动力学引擎，记录并推进世界的状态，比如物体质量、地面摩擦、飞行速度；一个图形引擎，把这个世界画成像素。 03:30



原课幻灯片：传统模拟器（状态动力学 + 图形引擎）对比从数据中学到的世界模型。

传统模拟器

- 有明确的 3D 场景和状态
- 可解释、可编辑、紧凑
- 通常能实时运行
- 不够真：画面和动力学都有仿真与现实的差距，比如复杂流体无法实时计算

世界模型

- 默认没有显式的 3D 状态
- 难以解释和编辑
- 实时可行，但大模型要跑到实时帧率并不容易
- 优势：画面逼真，物体和人的动作也更真实

三类用途

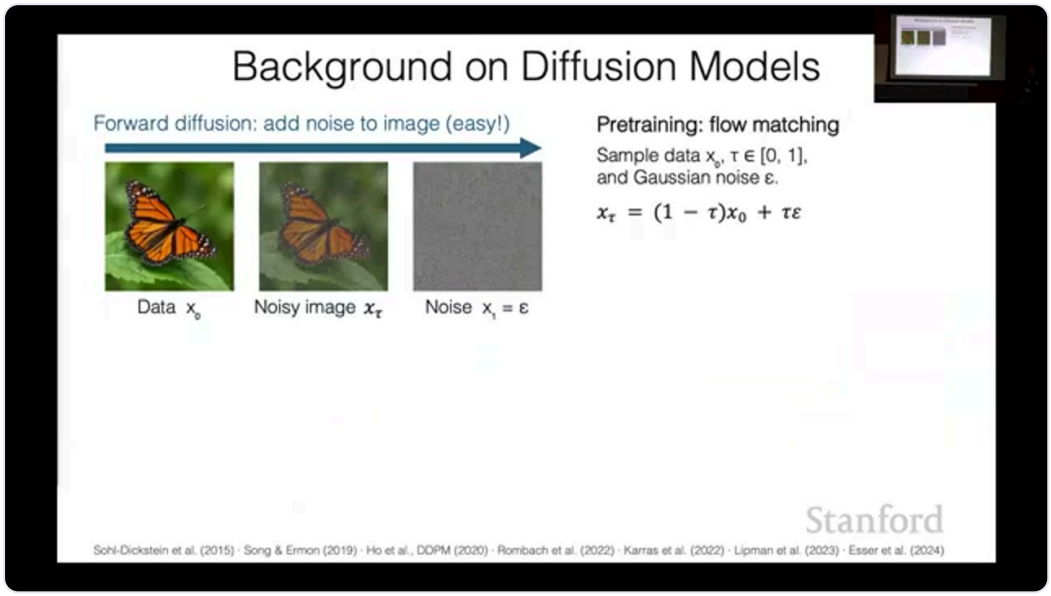
1. **训练和评估物理 AI**：让机器人、自动驾驶在里面交互。
2. **给人用**：游戏，或真实环境里练不起的培训，讲者举的例子是训练宇航员维修空间站。
3. **造数据**：特斯拉和 Waymo 都公开讲过，用世界模型生成没遇到过的罕见情况，比如马路中间出现一头大象，让驾驶策略更稳。 01:30–02:30

行业背景：讲者提到 Google DeepMind Genie 3、Waymo、特斯拉、World Labs、NVIDIA Cosmos，以及"几十家正在融资的创业公司"。(00:30–03:10)

PART 2

二 怎么造出来

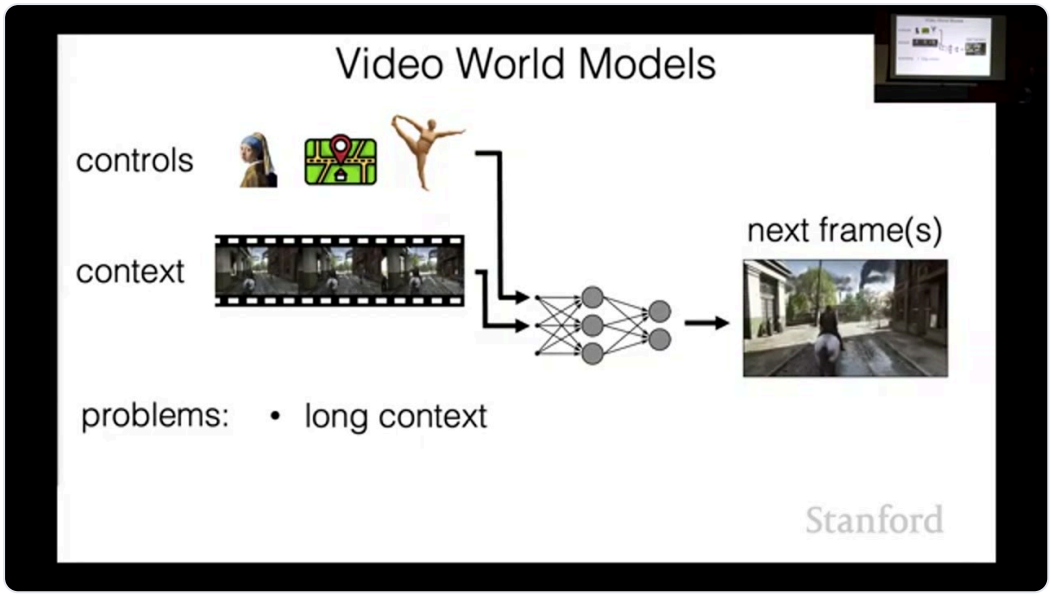
今天的世界模型大多建立在扩散模型（或其变体 flow matching）上，不是大语言模型那种"猜下一个 token"的自回归模型。07:00



原课幻灯片：前向扩散，把噪声逐步加到图像上；模型学习反过来去噪。

五个关键环节

1. **从噪声到数据**：直接从真实图像分布采样很难，于是从容易采样的高斯噪声出发，学习一条从噪声通往数据的轨迹。08:00
2. **训练方式**：给干净图像加噪（常用噪声和图像之间的线性插值），让网络预测"速度"，也就是沿轨迹变化的方向，损失函数是简单的 L2。10:30–11:30
3. **潜空间**：高分辨率下不直接处理像素，先用 VAE 压缩到潜空间，扩散过程在潜空间里进行。11:50
4. **网络结构**：大规模生成模型普遍使用 Transformer。文字描述经文本编码器后，通过交叉注意力注入。12:30–13:30
5. **数据规模**：学术界常用 ImageNet，一百多万张；商用的顶级图像生成模型，用的是几十亿张精心筛选的图像。09:00–10:00



原课幻灯片：视频世界模型 = 条件（controls）+ 上下文帧（context）→ 下一帧。

从图像到视频

视频世界模型在网络规模的视频上预训练：遮住几帧，用前面的帧去预测它们。推理时，前面几帧通过 KV 缓存作为上下文，和大模型根据前文预测下一个 token 是同一个思路。14:00–15:00

条件信号几乎可以是任何东西：人物身份照片、环境地图、动作捕捉、方向键、文字，甚至脑电、心跳、眼动、追踪到的手和头显位置。讲者把用于 VR 的这一类称作"生成式现实"。15:00–16:30

再往前一步是**世界动作模型（WAM）**：同时预测未来的画面和 Agent 未来的动作，机器人可以先"想象"未来，再从中得出动作。16:30

PART 3

三 三种用法

这是理解后面所有难题的骨架。区别在于：生成下一帧时，模型参考的是什么。17:00–19:30

开环 Open Loop	闭环 Closed Loop	锚定开环 Anchored
参考的是自己生成的帧。 典型场景：游戏、Genie 3。 问题：每个小错都会累积，产生漂移。	参考的是真实摄像头拍到的帧。 典型场景：机器人、自动驾驶。 每一步都被真实观察拉回，不容易漂。	自己生成画面，但有一个持续存在的锚，比如地图。 典型场景：讲者实验室的空间记忆、多人 Doom。 靠锚保持全局一致。

> 讲者列出的三大难题里，"记不住"和"越画越歪"主要发生在开环；"锚定开环"是他们针对前两个难题给出的共同思路；机器人则天然处在闭环里，这也是世界模型能先在机器人上落地的原因之一。

PART 4

四 三个难题

难题一：记不住（长上下文）

斯坦福一位本科生做的 Minecraft 世界模型 Oasis：玩家向前看、抬头看天、再低头，世界完全变了。因为上下文只装得下最近几帧，天空占满后，之前的画面就被挤出去了。19:00–20:30

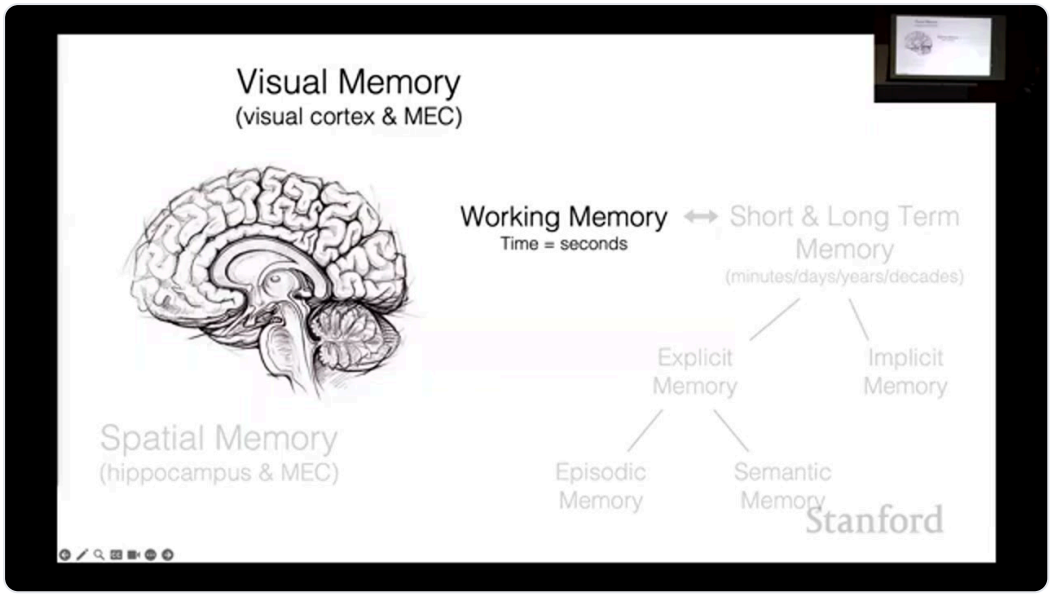


原课幻灯片：自回归逐帧生成的 Minecraft 世界模型。

为什么不能简单加长？Transformer 的自注意力，训练时计算量随 token 数平方增长，推理时线性增长。一段 512×512、每秒 32 帧、10 秒的视频，即使压缩到潜空间，也有几十万个 token。现在前沿大模型的上下文约一百万到两百万 token，放到视频上大约只多记一秒。21:00–22:40

借鉴人脑：人的记忆像一组专家系统，视觉记忆和空间记忆在不同脑区，工作记忆是秒级，短期和长期记忆是分钟到几十年。23:00–24:00

- **FramePack**：越久远的帧，在空间和时间上降采样得越多，全部打包成一张图，高效地交给扩散 Transformer。在 2025 年，这个方法就能让短视频里的人物面孔保持一致。24:00–25:30
- **空间记忆**：每帧额外生成深度图，融合成全局一致的点云；在当前位置渲染点云，作为下一帧的条件；同时保留几张关键帧。"走回胡佛塔，塔得还在，树叶变没变，没人在乎。"25:30–27:40

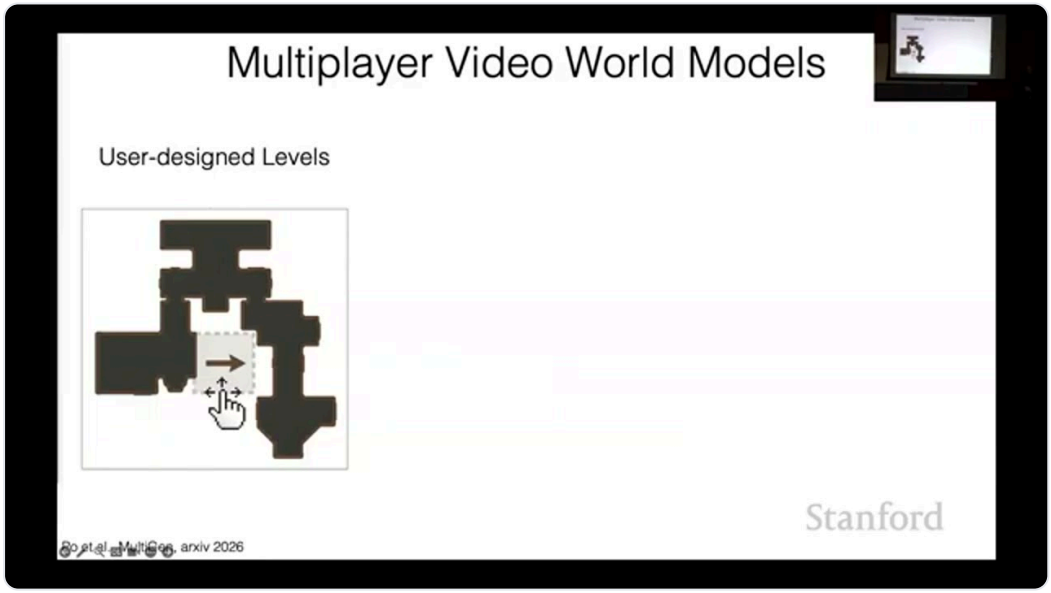


原课幻灯片：人脑中的视觉记忆、空间记忆与不同时间尺度的记忆系统。

难题二：多人共处一个世界

团队用 Doom 游戏引擎生成大量数据，做了一个多人对战的世界模型。方法是共享地图：把 2D 地图的墙体竖起来，渲染出玩家视角的深度图，作为条件，连同上下文帧和动作一起送进模型；玩家的位置、朝向、生死状态由一个单独的网络更新，存在地图上。27:40–30:30

- 只有几何结构是全局一致的，墙面纹理来自上下文窗口，未必每次一样，但玩家几乎察觉不到。
- 实验室成员连续玩了约 30 分钟，没有漂移；在 Modal 上公开托管一周，有几千名玩家参与。



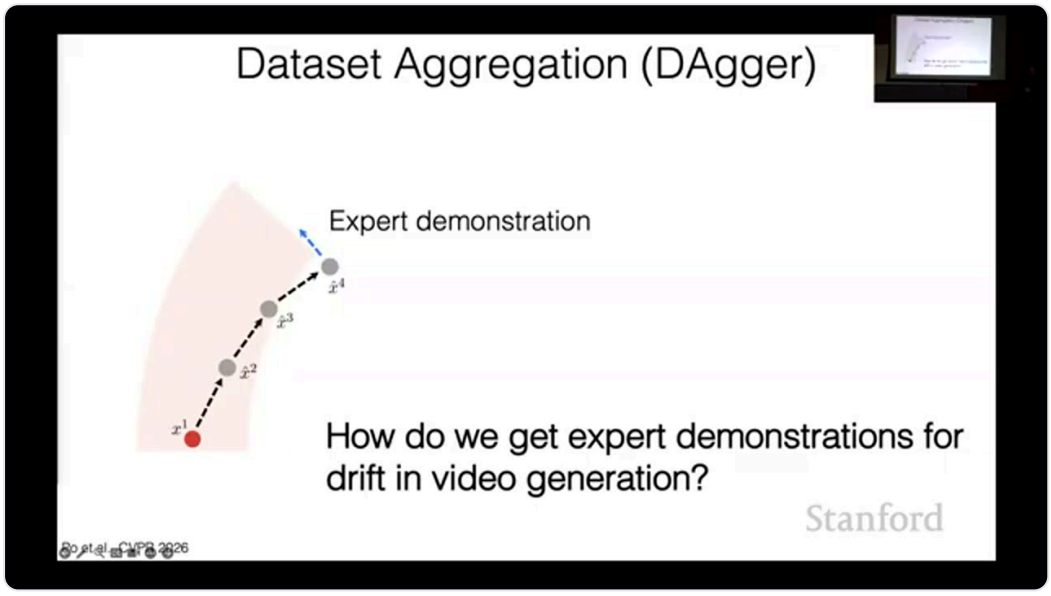
原课幻灯片：多人视频世界模型，所有玩家共享一张用户设计的关卡地图。

可控性

团队做过的控制方式：镜头控制（CameraCtrl 1/2）、用追踪到的手与生成世界中的物体互动（比如摸狗）、时空独立控制，也就是《黑客帝国》式的子弹时间，时间停住、镜头绕行。讲者提到，这些功能已经出现在 World Labs 等商用系统里。30:30–33:30

难题三：越画越歪（漂移）

自回归一直生成下去，动作变怪、颜色过饱和、物体变形。原因是训练时上下文都是干净的真实帧，推理时上下文是模型自己生成的、越来越差的帧，分布不一致，误差累积，即曝光偏差。33:10–34:20



原课幻灯片：DAgger，用专家示范帮助从偏离中恢复。问题是：视频生成里，专家示范从哪来?

BAgger（Backwards Aggregation）：让模型自己生成到出现退化，再把这段视频倒放，配上"这段是倒放"的文字提示（避免学到鱼倒着游这类不合理动作），就得到了从坏变好的纠正示范。多轮后训练后，模型学会自我纠正。对比：History Guidance 仍有明显颜色饱和；Self Forcing 画面好看但动作僵硬；BAgger 兼顾画面和动作。34:50–36:20

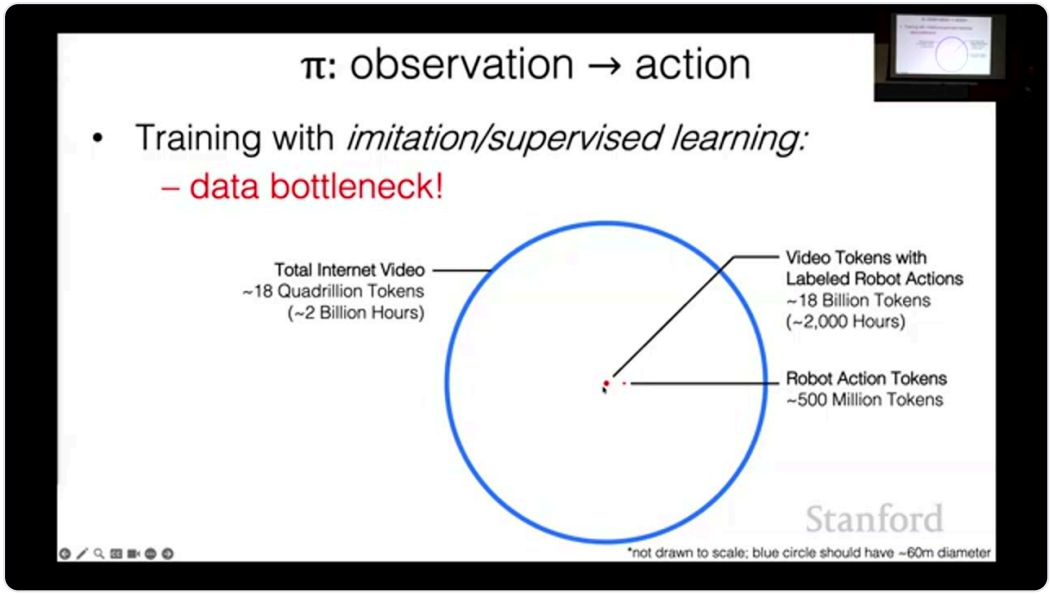
PART 5

五 从世界模型到世界智能体

驱动机器人的是策略：从观察到动作的映射，闭合感知—动作循环。讲者比较了三条训练路线：37:00–39:00

路线	优点	问题
传统模拟器	数据几乎免费、无限	仿真和现实的差距，数据多样性永远比不上真实世界

完美世界模型当模拟器	原理上数据多且没有差距	这样的模型还不存在，机器人很容易把它"玩坏"；可能需要和传统模拟器混合
模仿学习	当前主流，如 diffusion policy	带动作标注的数据太少



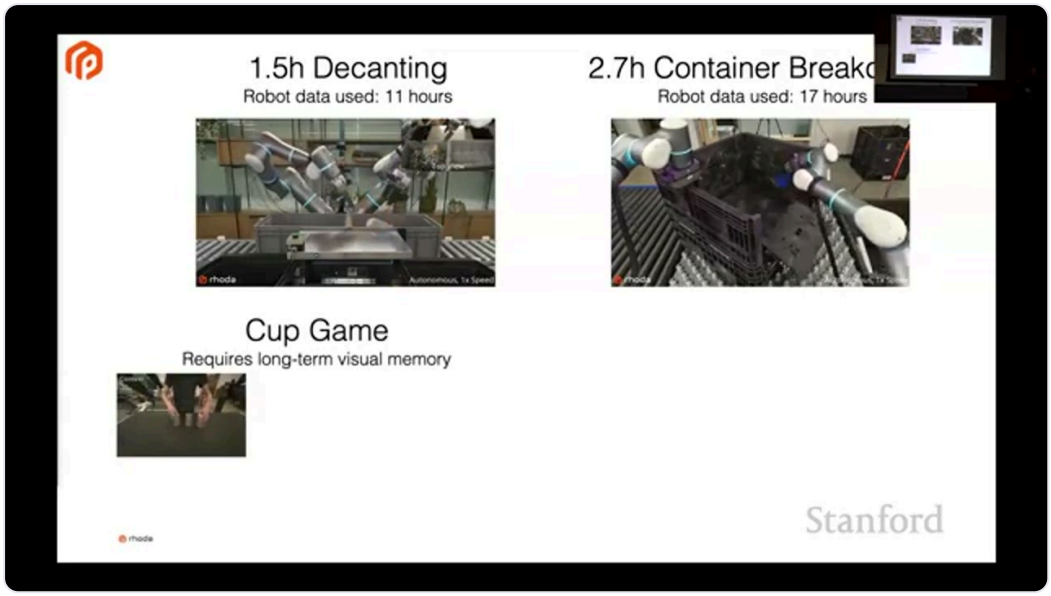
原课幻灯片：全网视频约 18 千万亿 token（约 20 亿小时）；带机器人动作标注的视频约 180 亿 token（约 2000 小时）；机器人动作 token 约 5 亿。注：若按比例画，蓝色圆的直径约 60 米。

为什么数据量是关键：2020 年的大模型缩放定律论文表明，思维链、上下文学习这类涌现能力，只在足够大的模型和数据上出现。机器人要么把"小红点"做大，很多公司正在大量采集机器人数据；要么把海量无标注视频用起来。 40:30–41:20

Rhoda 的做法：先想象，再反推

1. **视频生成**：策略把当前观察映射为"未来的观察"，即一段机器人成功完成任务的视频。这一步可以用网络规模的无标注视频训练。
2. **逆动力学**：从想象出的视频反推末端执行器的位姿和动作。这一步数据效率很高，不需要多少数据。

讲者说明：这部分是他参与的创业公司 Rhoda 的工作，不是斯坦福实验室的项目；他们约两年前开始做基于视频的机器人，现在这个方向已经很拥挤。（37:00–45:00）



原课幻灯片：开箱倒料（Decanting）、拆箱（Container Breakdown）、需要长期视觉记忆的猜杯子任务。幻灯片标注：开箱倒料 1.5h，所用机器人数据 11 小时；拆箱 2.7h，所用机器人数据 17 小时。

已经在做的工业任务

- **开箱倒料**：大型汽车制造商定义的任务。机械臂自主打开沉重的箱子，把轴承倒进周转箱，把塑料件和蜡纸分别放进回收箱。
- **拆箱**：农业场景里装农产品的箱子，用完后需要清理、拆解。
- **退货检查**：检查退回的衣服并重新打包；记住哪些已经检查过，效率会高很多。
- **猜杯子**：物体藏在杯下来回换位，只有上下文足够长才能解，演示里机器人先生成"成功完成"的视频，再执行。

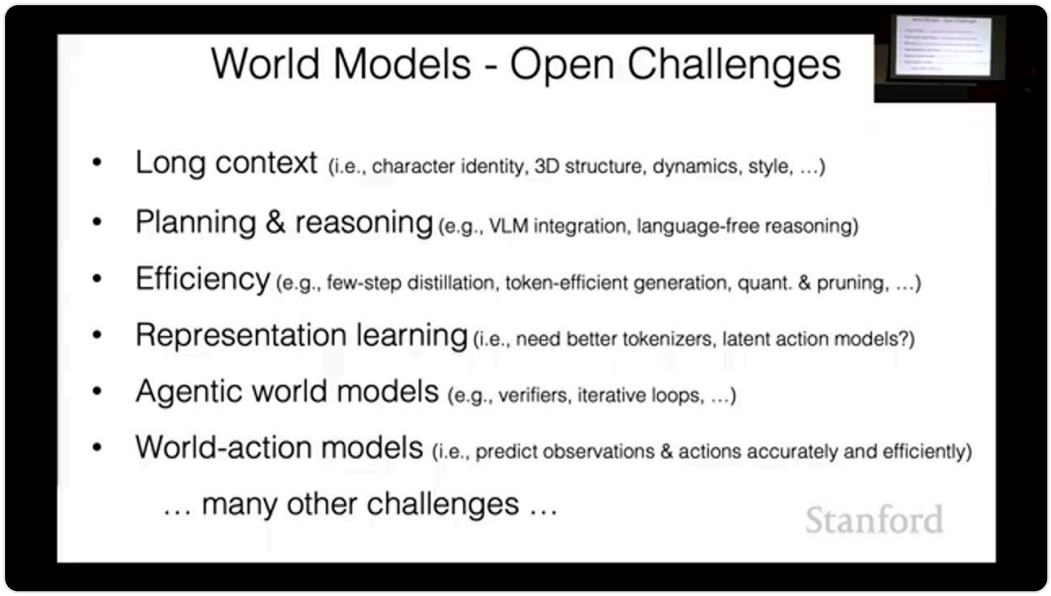
（42:40—44:30）

安全与幻觉

- 闭环：每次只往前想一小段（大约一秒、几帧），执行，再根据真实观察继续。49:00
- 检验：用视觉语言模型或感知模型当检验员，判断结果是否符合目标；也可以先生成多个可能的未来，挑一个再执行。49:00—50:00
- 分工：长时程的任务规划交给视觉语言模型，视频模型负责下一段物理动作规划；触觉、声音等信号也有用。53:30—54:10
- "任何生成模型都会幻觉。"关键是贴着物理观察，而不是规划几分钟、几小时以后的事。53:00

PART 6

六 还没解决的问题



原课幻灯片：世界模型的开放挑战。

1. **长上下文**：永远不够长。要记住 3D 结构、做过的事，也是外推动力学、上下文学习的基础。
2. **规划与推理**：怎么和视觉语言模型结合最好？推理放在视觉语言模型里，还是放在扩散模型里？
3. **效率**：少步蒸馏（30 – 50 步压到 1 – 2 步）、只关注上下文里真正重要的 token、量化、剪枝、模型架构。
4. **表示学习**：现在的 VAE 不够好，还没有可扩展的、专为生成任务设计的表示学习方法。
5. **Agent 化的世界模型**：用 GPT-6 这类会用工具、能闭环的 Agent 当检验员，或放进迭代循环。
6. **世界动作模型**：怎样最好地同时生成动作和观察？推理时是否一定需要观察？

(44:40–46:40)

问答精选

时间怎么表示？默认用帧序号当时间，但如果要独立控制场景的全局时间和镜头，就不够了。归根结底是数据问题：需要大量各个变量独立变化的训练样本。47:30–48:40

模型怎么知道自己错了？闭环里只想一小段，执行后用检验员对照目标；也可以先模拟多个未来，再挑一个。49:00–50:00

开源基础模型够用吗？预训练极难：架构、海量数据、标注、巨量算力缺一不可。一个 140 亿参数的扩散 Transformer，怎么蒸馏也很难跑到实时，延迟和成本都高。开源模型进步很大，但公司越来越不愿意开源，现有模型还不够好。50:00–52:00

点云地图越来越大怎么办？放进上下文的是从当前视角渲染出的点云图，本身不占多少空间；真正的问题是不断增长的地图怎么存，可以考虑高斯、网格或 2D 地图。51:30–52:40

SOURCES

原课与延伸

- 原课视频：Stanford EE292R AI and Data for Science · Autumn 2026 · From World Models to World Agents（Stanford Online，YouTube：youtube.com/watch?v=IfkeGU4bEJI）
- 课程日程与摘要：aida4sci.stanford.edu/schedule.html
- 讲者实验室：psAI.stanford.edu（斯坦福物理与空间 AI 实验室）
- 文中时间点均对应原课视频，格式为"分:秒"。

说明：本笔记为个人学习整理，原课为英文，中文为整理者翻译和概括，如有出入以原课为准。文中幻灯片截图版权归原作者和斯坦福大学所有，仅用于学习交流。"一个整理者的观察"为整理者个人看法，不代表讲者观点。

心晴九天

持续拆解硅谷一手课程与 AI Agent 落地。欢迎转发给需要的朋友。