

AMCC 5160 · Lecture 09 transcript

Worlds You Can Walk Into, Robots That Learn by Watching • 27 October 2026

1. Worlds You Can Walk Into, Robots That Learn by Watching

Good evening, everyone, and welcome to Week 9. For eight weeks we have been making pictures that move: images, motion, long takes, and last week, video that talks. Tonight the picture stops being a picture. It becomes a world you can walk into, a 3D place you can export to Blender, and finally a body: a robot that learned what to do by watching people. Three words for tonight: worlds, splats and bodies.

欢迎来到第九讲。八周以来我们一直在做会动的画面：图像、动作、长镜头，上周还有会说话的视频。今晚，画面不再只是画面：它变成可以走进的世界，变成可以导出到 Blender 的 3D 空间，最后变成一个身体——一个通过看人类视频学会做事的机器人。今晚的三个关键词：世界、高斯点（splats）、身体。

2. I push this cup off the edge of the table. What happens next?

Let's start with a tiny game. [Hold a cup near the edge of the desk.] If I push this cup off the edge, what happens next? Shout it out. Where does it land, what sound, does it break? Now, what if the floor were carpet? What if I pushed it gently? Notice what you just did. You predicted the future of a scene, and you changed your prediction when I changed the action. That is exactly what tonight's models try to learn: what happens next, given what I do.

先做个小游戏。我把杯子推下桌子，会发生什么？落在哪里、什么声音、会不会碎？如果地上是地毯呢？如果我轻轻推呢？你们刚才做的事情，就是在脑子里预测场景的未来，而且当我改变动作时，你们的预测也改变了。这正是今晚的模型要学的：在我做某个动作之后，世界会怎样变化。

3. Four stories from this year

Four stories tonight, same format as the last two weeks. One: video that you can steer, frame by frame, which is what people now call an interactive world model. Two: 3D comes back, and generated worlds become Gaussian splats you can export, with one of the biggest deals of the year behind it. Three: robots that learn by watching people, including a model released yesterday. Four: a real disagreement in the field. Does a world model need to draw pixels at all? We'll finish with a debate.

今晚四个故事，和前两周的格式一样。一：可以一帧一帧操控的视频，也就是交互式世界模型。二：3D 回归，生成的世界变成可以导出的高斯点云，背后还有今年最大的交易之一。三：通过看人学习的机器人，包括昨天

刚发布的模型。四：领域里的真实分歧——世界模型一定要生成像素吗？最后我们辩论。

4. Where we've been: the pieces you already have

You already own most of the pieces. In Lecture 03 we asked what a world model predicts, we did the round-trip test with Astra, and we saw a vision-language-action model for the first time. In Lecture 07, Helios generated a minute of video in real time using a KV cache and distillation, and a humanoid learned a K-pop dance from phone video. Last week, real-time avatars that listen. Tonight we add one thing to each: an action in the loop, explicit geometry, and a real body.

大部分零件你们已经有了。第三讲我们问过世界模型在预测什么，做过 Astra 的“往返测试”，第一次见到 VLA。第七讲 Helios 用 KV 缓存和蒸馏实时生成一分钟视频，人形机器人从手机视频学会了 K-pop 舞。上周是会倾听的实时虚拟人。今晚每样东西再加一步：把动作放进循环、保留几何、接上真实的身体。

5. You have seen two slides of tonight already

Here are two slides from Lecture 03. On the left, the question 'what is being predicted?': a pixel, a compact state, or an action's result. On the right, the VLA: the robot sees the scene, reads the task and acts. In September these were one slide each, near the end of a hundred-slide lecture. Tonight each of them becomes a whole story. If you have your Lecture 03 notes, keep them open.

这是第三讲的两页幻灯片。左边：世界模型“在预测什么”——像素、紧凑状态，还是动作的结果？右边：VLA，机器人看场景、读任务、执行动作。九月份它们各只有一页，今晚各自展开成一个完整的故事。

6. How a top course maps the field: three paradigms

This is how Matt Gormley and Aran Nayebi at CMU organise their world-models lecture, which we also borrowed from in Lecture 03. Three paradigms. First, generate a 3D scene, then render it, like World Labs' Marble. Second, interactive video, like DeepMind's Genie: you press a key and the next frame appears. Third, a latent world representation, where you predict an abstract state instead of pixels, like Meta's V-JEPA. Our four stories follow this map, and robots borrow from all three.

这是 CMU 的 Matt Gormley 和 Aran Nayebi 讲世界模型的框架，第三讲我们也借用过。三种范式：一、先生成 3D 场景再渲染，比如 World Labs 的 Marble；二、交互式视频，比如 DeepMind 的 Genie，按一个键，下一帧出现；三、隐空间的世界表示，预测抽象状态而不是像素，比如 Meta 的 V-JEPA。今晚四个故事就按这张地图走，机器人则三者都用。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley

7. From video to playable worlds

Story one. A video model gives you a clip. A world model gives you a place: you press a key, or type an event, and it draws the next moment for you, live.

第一个故事。视频模型给你一段片子；世界模型给你一个地方：你按一个键、或输入一个事件，它就实时画出下一刻。

8. What is a world model?

CMU's definition is simple: a world model takes the previous state of the world and an action, and predicts the next state, or a distribution over next states. The state might be a picture, a 3D scene or an abstract vector. The important word is action. A video model answers 'what happens next?' A world model answers 'what happens next if I do this?' That one extra input changes the medium from film to game.

CMU 的定义很简单：世界模型输入上一时刻的世界状态和一个动作，预测下一时刻的状态（或状态分布）。状态可以是图像、3D 场景或抽象向量。关键词是“动作”。视频模型回答“接下来会怎样”，世界模型回答“如果我这样做，接下来会怎样”。多了这一个输入，媒介就从电影变成了游戏。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

9. Eight years, from a dream of a car race to a playable world

A quick history. In 2018, David Ha and Jürgen Schmidhuber trained an agent to drive inside its own dream of a car-racing game, and called the paper World Models. In 2024 a Google team ran DOOM entirely inside a diffusion model, GameNGen, at about twenty frames per second. The same year, DeepMind's first Genie learned playable 2D platform games from internet video, at about one frame per second. Genie 2 went 3D. Genie 3, August 2025, reached real time: 720p at 24 frames per second, consistent for a few minutes. And in 2026, the public could try it, and Google started calling a product a world model.

简短历史。2018 年 Ha 和 Schmidhuber 让智能体在它自己“梦到”的赛车游戏里学开车，论文就叫《World Models》。2024 年谷歌团队用扩散模型直接“运行”DOOM，GameNGen，约 20 帧每秒。同年 DeepMind 的第一代 Genie 从网络视频学会可玩的 2D 平台游戏，约 1 帧每秒。Genie 2 进入 3D。2025 年 8 月的 Genie 3

达到实时：720p、24 帧、保持几分钟一致。2026 年公众可以试用，谷歌也开始把一个产品叫做“世界模型”。

10. Two ancestors: a dream of a race, and DOOM without an engine

Two ancestors worth seeing. On the left, Ha and Schmidhuber's car racer. One half is a real frame from the game; the other is the same frame as the agent's compressed model remembers it. The agent learned to drive partly inside this blurry dream. On the right, GameNGen. There is no game engine here at all: every frame of DOOM is generated by a diffusion model reacting to the player's keys. The left idea is 'practise inside a model'; the right idea is 'the model is the game'. Genie joins the two.

两个值得一看的祖先。左边是 Ha 和 Schmidhuber 的赛车：一半是真实游戏画面，一半是智能体压缩模型记住的样子，它部分地在这个模糊的梦里学会了开车。右边是 GameNGen：根本没有游戏引擎，DOOM 的每一帧都由扩散模型根据玩家按键实时生成。左边的想法是“在模型里练习”，右边是“模型就是游戏”。Genie 把两者结合了。

11. Genie 3: a world that waits for you

Genie 3. [Play.] You type a prompt, a world appears, and you walk around in it with the keyboard. It runs at 720p and 24 frames per second, in real time. If you look away and come back, the wall you painted is still there, for about a minute. And there is a new kind of control DeepMind calls promptable world events: while you are walking, you type 'a storm starts', or 'a deer runs across', and the world changes. For a director, that is a new verb.

Genie 3。输入提示词，世界出现，你用键盘在里面走。720p、24 帧、实时。你转头再转回来，刚才涂的墙还在，大约能记住一分钟。还有一种新的控制，DeepMind 叫“可提示的世界事件”：走着走着输入“暴风雨开始了”或“一只鹿跑过”，世界就改变。对导演来说，这是一个新动词。

12. Project Genie: anyone (with a subscription) can try it

In 2026 Google turned Genie 3 into a public prototype called Project Genie. It combines Genie 3 with the Nano Banana Pro image model and Gemini. Notice the limits: US Ultra subscribers only, and each world lasts sixty seconds. That cap tells you what these things cost to run: every frame is generated, live, for one person. People immediately made imitations of famous Nintendo games, and video-game company stocks dipped. Hold that thought for the debate.

2026 年谷歌把 Genie 3 做成了公开原型 Project Genie，结合了 Genie 3、Nano Banana Pro 图像模型和 Gemini。注意限制：只对美国 Ultra 订阅用户开放，每个世界 60 秒。这个上限说明运行成本有多高：每一帧都是为一个人实时生成的。大家马上用它模仿著名的任天堂游戏，游戏公司股价还跌了一下。辩论时记住这一点。

13. Waymo World Model: Genie 3 learns to drive

Here is the same model doing serious work. Waymo built a variant of Genie 3 for self-driving. It generates not only camera views but lidar, the laser depth sensor on the car's roof. Engineers can control the car's actions, the layout of the street, what other drivers do, and the weather. Why? Because the dangerous situations you most need to train on, a child running out between parked cars, are the ones you can never film on purpose. A world model is a place to rehearse the rare.

同一个模型在做严肃的工作。Waymo 基于 Genie 3 做了自动驾驶版本，不仅生成摄像头画面，还生成激光雷达数据。工程师可以控制车辆动作、街道布局、其他司机的行为和天气。为什么？因为最需要训练的危险情况，比如孩子从停着的车之间跑出来，你永远不能故意去拍。世界模型是排练罕见情况的地方。

14. Gemini Omni: “world model” becomes a product name

In May, Google launched Gemini Omni and called it a world model in the keynote. In practice, most people meet it as a video generator and editor: you give it text, audio, images or video, and you change the camera, background or style by talking to it. Every output carries a SynthID watermark. I show this to make a point about words: 'world model' is now also marketing. Part of tonight is learning to ask, which kind?

五月谷歌发布 Gemini Omni，在主题演讲里称它为世界模型。实际上大多数人把它当视频生成和编辑工具用：输入文字、音频、图片或视频，用对话改镜头、背景、风格。所有输出都带 SynthID 水印。我放这个是想说明：“世界模型”现在也是营销词。今晚要学会问一句：是哪一种世界模型？

15. Under the hood: one extra input

Under the hood, the difference is one input. A video model predicts the next frame from the past frames and a text prompt. A world model also takes your action at every step: a key press, a joystick, or a typed event. The network can be almost the same autoregressive video model we met with Helios in Lecture 07. But because the action arrives every step, the model can't generate the whole clip in advance; it must wait for you and answer quickly.

原理上，区别只是多一个输入。视频模型根据过去的帧和文字预测下一帧；世界模型每一步还要接收你的动作：按键、摇杆或输入的事件。网络可以几乎就是第七讲 Helios 那种自回归视频模型。但因为每一步都有新动作进来，模型不能提前生成整段视频，它必须等你，并且快速回应。

16. Under the hood: where do the actions come from?

Here's a real problem. To learn what happens if I press left, you need video labelled with key presses,

and the internet doesn't have that. Genie's trick, shown in this CMU slide: a latent action model looks at two consecutive frames and asks, what small action could explain the change? It must squeeze the answer into one of only eight codes. Nobody labels anything. After training, the codes turn out to mean things like left, right, jump. At play time, your key press simply chooses a code.

这里有个实际问题：要学“按左键会怎样”，需要标注了按键的视频，但网上的视频没有这种标注。Genie 的办法：潜在动作模型看相邻两帧，问“什么小动作能解释这个变化？”答案只能压缩成 8 个代码之一。没有任何人工标注。训练完后，这些代码自然对应左、右、跳等。玩的时候，你的按键就是选一个代码。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

17. Genie at play time: three parts in a loop

At play time, Genie is a loop of three parts. A video tokenizer turns the frame into tokens, like the compressed representation in Lecture 03. Your key press becomes a latent action. A dynamics model, a transformer, predicts the next frame's tokens from the past tokens and the action. Decode, show it to you, repeat. Twenty-four times a second, in Genie 3's case.

玩的时候 Genie 是三部分的循环：视频分词器把画面变成 token，就像第三讲讲的压缩表示；你的按键变成潜在动作；动力学模型（一个 Transformer）根据过去的 token 和动作预测下一帧的 token。解码、显示、重复。Genie 3 每秒 24 次。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

18. Be the latent action model

Let's be the latent action model. I'll show two frames from a game you've never seen. Each row has only eight hand signals to explain the change between them. Agree quickly as a row. [Show frames from the Genie qualitative results slide, or any two frames from a platform game.] Now compare rows. Did you all invent the same actions? Probably left, right and jump appear in every row, but the rest differ. That's exactly what happens in the model: the codes are discovered, not given, so they are useful but not always tidy.

我们来当潜在动作模型。我给你们看一个没见过的游戏的两帧画面，每一排只有 8 个手势来解释两帧之间的变化，快速商量。现在比较各排：大家发明的动作一样吗？左、右、跳大概每排都有，其他的就不同了。模型里也是这样：动作代码是“发现”的而不是给定的，有用但不一定整齐。

19. Under the hood: why the world must come back the same

Remember the round-trip test from Lecture 03? See a red door, turn away for ten seconds, turn back. A good world model shows the same red door. A weak one shows a blue door, or no door, because it only remembers what is in its context window. It's the same drift problem we saw in Lecture 07's long takes. Genie 3 holds visual memory for about a minute. That's the main thing to test when you try any world model: walk away, and come back.

还记得第三讲的往返测试吗？看到一扇红门，转开十秒，再转回来。好的世界模型显示同一扇红门；差的会变成蓝门，或者门没了，因为它只记得上下文窗口里的东西。这和第七讲长镜头的漂移是同一个问题。Genie 3 的视觉记忆大约一分钟。测试任何世界模型最重要的一招：走开，再回来。

20. Under the hood: the real-time budget

The budget. At 24 frames per second you have 41.7 milliseconds per frame to read the action, run the model, decode and display. You already know the tricks: a KV cache so you don't recompute the past, and distillation from fifty denoising steps to a few. Those made Helios real-time in Lecture 07 and the avatars real-time last week. A world model is the same budget, except now your hand is on the controls.

时间预算：24 帧每秒，每帧只有 41.7 毫秒，要读取动作、运行模型、解码、显示。技巧你们都学过：KV 缓存避免重算过去，蒸馏把 50 步去噪压到几步。第七讲的 Helios、上周的虚拟人都是这样做到实时的。世界模型是同样的预算，只是现在你的手在控制。

21. Genie 3 vs. Genie 2 vs. GameNGen

CMU's comparison table. Read it left to right. GameNGen: one game, DOOM. Genie 2: general 3D worlds, but short and not real time. Genie 3: 720p, general, minutes long, real time. The biggest jump isn't image quality; it's the interaction horizon, from seconds to minutes. That's what turns a demo into a place you can spend time in.

CMU 的对比表，从左往右读。GameNGen 只有一个游戏 DOOM；Genie 2 是通用 3D 世界但很短、不实时；Genie 3 是 720p、通用、几分钟、实时。最大的跃升不是画质，而是交互时长：从几秒到几分钟。这让演示变成了你能待在里面的地方。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

22. What world models still get wrong

And the limits, from the same lecture. Minutes, not hours. A limited, fixed set of actions. Multiple characters interacting is hard. Text and signs in the world are unreliable. And the last point leads to Story 2: it's video. There's no mesh, no 3D file. When the session ends, the world is gone. If you want to keep the world, open it in Blender, or walk a robot through it, you need geometry.

局限，来自同一讲：几分钟而不是几小时；动作集合有限且固定；多个角色互动很难；世界里的文字不可靠。最后一点通向第二个故事：它是视频，没有网格，没有 3D 文件。会话结束，世界就消失。想保存世界、在 Blender 里打开、让机器人在里面走，就需要几何。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

23. A bridge to robots: Genie dreams of a robot arm

One more slide from CMU, as a bridge. Already in 2024 the Genie team trained the same recipe on robot videos with no action labels, and the latent actions became arm movements. Keep that in mind: if a model can imagine what happens when an arm moves, a robot can practise inside that imagination. That's Story 3.

CMU 的最后一页，作为桥梁。2024 年 Genie 团队就把同样的方法用在无动作标注的机器人视频上，潜在动作变成了机械臂的动作。记住：如果模型能想象手臂动了之后会怎样，机器人就能在这种想象里练习。这是第三个故事。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

24. Useful: directing a world, not a clip

Something useful for your projects. When you direct a world instead of a clip, think in four parts. The world: place, era, weather, light. The body: who you are, how tall, how fast, which camera. The events you can trigger while you walk. And the test: walk away and come back, and try to break it. For Assignment 3, your game, a world model is a fast way to sketch a level and feel it, before you build the parts that work in Unity or the browser.

给项目的实用建议：导演一个世界而不是一段片子，分四部分想。世界：地点、年代、天气、光线。身体：你是谁、多高、多快、什么镜头。事件：行走中可以触发什么。测试：走开再回来，试着把它弄坏。对作业三的游戏来说，世界模型是快速草拟关卡、先感受一下的方法，然后再在 Unity 或浏览器里把好的部分做出来。

25. If a world only exists while someone is playing it, what is the artwork: the world, the play, or the recording?

Turn to your neighbour for three minutes. If a world only exists while someone is playing it, and disappears when the session ends, what is the artwork? The model? The prompt? Your sixty seconds of play? Or the screen recording you post afterwards?

和旁边的同学讨论三分钟：如果一个世界只在有人玩的时候存在，会话结束就消失，那艺术作品是什么？模型？提示词？你那 60 秒的游玩？还是你之后发的录屏？

26. 3D is back: worlds as Gaussian splats

Story two. If video worlds disappear, the other answer is: generate the world in 3D, so you can keep it.

第二个故事。如果视频世界会消失，另一种答案是：直接生成 3D 世界，这样就能保存下来。

27. Pixels or geometry? Two ways to make a world

Two ways to make a world. Draw the next frame, like Genie: anything can happen, water, crowds, storms, and it feels alive, but it forgets, and there's nothing to export. Or build the scene first and render it, like Marble, Atlas or HY-World: it's a file, consistent forever, you can open it in Blender, Unity, a VR headset or a robot simulator. But it's mostly static. Both camps are moving towards each other, and Atlas, which we'll see in a minute, is an attempt to have both.

造世界的两种方式。逐帧画，像 Genie：什么都能发生，水、人群、暴风雨，很鲜活，但会遗忘，也没有东西能导出。或者先建场景再渲染，像 Marble、Atlas、HY-World：它是一个文件，永远一致，能在 Blender、Unity、VR 头显或机器人模拟器里打开，但大多是静态的。两派在互相靠近，等一下要看的 Atlas 就是想两者兼得。

28. Marble: one picture in, a world you can walk around

Marble, from Fei-Fei Li's company World Labs. You give it text, a photo, several photos, a 360 panorama, or a video, and it gives you a world you can walk around. The output isn't video; it's a Gaussian splat file, .spz or .ply, which you can download. Version 1.1 in April improved lighting and added worlds that expand automatically. You can view it on a laptop, a phone or a VR headset with their open-source renderer, Spark.

Marble，来自李飞飞的公司 World Labs。给它文字、一张照片、几张照片、360 全景或视频，它给你一个可以四处走动的世界。输出不是视频，而是高斯点文件（.spz 或 .ply），可以下载。四月的 1.1 版改进了光照，世界

可以自动扩展。用开源渲染器 Spark，可以在电脑、手机或 VR 头显上看。

29. Atlas (1 Sept 2026): one model for pictures, video and 3D

In September World Labs announced Atlas, the model behind future versions of Marble. The idea: every input, text, photos, depth, video, goes into one shared spatial context, and each image is pinned to a camera position in 3D. From that context it can generate new views, up to a minute of 1440p camera-controlled video, depth maps, point clouds or splats. With only two or three ordinary phone photos it can rebuild a scene faithfully. It's in early access only.

九月 World Labs 发布 Atlas，未来 Marble 版本背后的模型。思路是：所有输入——文字、照片、深度、视频——都放进同一个“空间上下文”，每张图片都固定在 3D 中的一个相机位置。由此它能生成新视角、最长一分钟 1440p 的相机控制视频、深度图、点云或高斯点。只用两三张普通手机照片，就能忠实重建场景。目前只开放给部分合作者。

30. Atlas, by World Labs' own numbers

World Labs' own numbers. Three quarters or more of raters preferred Atlas for camera-controlled video, against five video models including Seedance 2.5, which we saw in Lecture 07. Two or three photos for a faithful reconstruction. Twenty-four phone-video frames per room to simulate a robot driving through it, which links to Story 3. And a lower 3D reconstruction error than specialised open models. But these are company numbers, with no paper. Treat them as claims to test.

World Labs 自己公布的数据：在相机控制视频上，75%到 94%的评分者更喜欢 Atlas，对手包括第七讲见过的 Seedance 2.5。两三张照片就够重建。每个房间 24 帧手机视频就能模拟机器人在里面行驶，这连到第三个故事。3D 重建误差也低于最好的开源模型。但这些都是公司数据，没有论文，把它们当作待验证的说法。

31. The money: a chip company buys a world-model company

Now the money. In February, World Labs raised a billion dollars, from AMD, NVIDIA and Autodesk among others. At the end of September, AMD announced it would buy World Labs in an all-stock deal worth about 8.2 billion dollars, with Fei-Fei Li becoming AMD's chief scientist. Why would a chip maker buy a world-model company? Because worlds are compute: every generated frame, and every robot practising inside a simulated world, runs on chips. The deal hasn't closed, and nobody has said what happens to Marble or Spark.

再看钱。二月 World Labs 融资 10 亿美元，投资方包括 AMD、NVIDIA 和 Autodesk。九月底 AMD 宣布以约 82 亿美元全股票交易收购 World Labs，李飞飞将成为 AMD 首席科学家。芯片公司为什么买世界模型公司？因为

世界就是算力：每一帧、每个在模拟世界里练习的机器人都跑在芯片上。交易尚未完成，Marble 和 Spark 的未来也没有公布。

32. The open alternative: Tencent HY-World 2.0

If you want something open, Tencent's HY-World 2.0. From text or one image it builds a navigable 3D world, and from photos or video it reconstructs one. The pipeline is easy to read: make a 360 panorama, plan a camera path, expand the views, and then one network, WorldMirror 2.0, predicts depth, normals, cameras and Gaussian splats in one pass. It exports to Blender, Unity, Unreal, and NVIDIA's Isaac Sim for robots. Code and weights came out in April and May.

想要开源的，可以用腾讯混元的 HY-World 2.0。文字或一张图生成可导航的 3D 世界，照片或视频则做重建。流程很好理解：先生成 360 全景，规划相机路径，扩展视角，然后 WorldMirror 2.0 一次性预测深度、法线、相机和高斯点。可以导出到 Blender、Unity、Unreal 和机器人用的 Isaac Sim。四五月份开源了代码和权重。

33. Under the hood: from a scene to a function (NeRF)

How does a pile of photos become a 3D world? Stanford's CS231A tells the story in two steps. Step one, NeRF, 2020: a neural network that answers a question. At this point in space, seen from this direction, what colour is it and how dense? To render a pixel, you march along the camera ray and ask the network many times. You train it until its renders match your photos. Beautiful results, but slow, because you ask the network millions of times per image.

一堆照片如何变成 3D 世界？斯坦福 CS231A 分两步讲。第一步，2020 年的 NeRF：一个神经网络回答问题——空间中这个点，从这个方向看，是什么颜色、多大密度？渲染一个像素时，沿相机光线走，问网络很多次。训练到渲染结果和你的照片一致。效果漂亮，但很慢，每张图要问网络上百万次。

Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

34. Under the hood: only put stuff where stuff is

Step two, 3D Gaussian Splatting, 2023. The key idea on this slide: most of space is empty air, so why ask a network about it? Instead, put millions of small, soft, coloured blobs, 3D Gaussians, only where something exists. Each blob has a position, a shape, an opacity and a colour. To render, you throw each blob onto the screen, like a splat of paint. That's the name.

第二步，2023 年的 3D 高斯泼溅。这页的关键想法：大部分空间是空气，为什么还要问网络？改成只有在有东西

的地方放上百万个柔和的彩色小团，也就是 3D 高斯。每个高斯有位置、形状、不透明度和颜色。渲染时把每个高斯“泼”到屏幕上，像一团颜料，名字就是这么来的。

Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

35. Worked example: what is inside one splat?

Let's count. One splat has three numbers for position, seven for shape, three scales and a four-number rotation, one for opacity, and forty-eight for colour, because the colour changes with viewing angle and is stored as spherical harmonics, sixteen coefficients for each of red, green and blue. Fifty-nine numbers, about 236 bytes. A million splats is about 236 megabytes raw. A room has one to three million. Compressed formats like .spz, which Marble uses, are roughly ten times smaller. That's why splats now stream to a phone.

算一算。一个高斯有 3 个位置数、7 个形状数（3 个缩放加 4 个旋转）、1 个不透明度、48 个颜色数——颜色随视角变化，用球谐函数存储，红绿蓝各 16 个系数。共 59 个数，约 236 字节。一百万个约 236MB。一个房间一到三百万个。Marble 用的.spz 压缩格式大约小十倍，所以现在手机也能流式加载。

36. Under the hood: rasterize, don't ray-march

Why is it fast? Instead of marching along rays and asking a network, you project every blob onto the screen as a flat ellipse, sort them by depth, and blend them front to back. That's rasterization, the same thing graphics cards do for video games. At render time there is no neural network at all, just millions of coloured ellipses.

为什么快？不再沿光线问网络，而是把每个高斯投影到屏幕上变成 2D 椭圆，按深度排序，从前往后混合。这就是光栅化，显卡给游戏做的事情。渲染时完全不需要神经网络，只有上百万个彩色椭圆。

Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

37. The result: from 0.07 to 100+ frames per second

And here is the payoff. High-quality NeRF rendered at about 0.07 frames per second. Instant-NGP got to about nine. Gaussian splatting: over a hundred. Real time is what turned splats from a paper into a medium you can walk around in VR. Same lesson as Lecture 07: speed changes what something is.

结果：高质量 NeRF 大约每秒 0.07 帧，Instant-NGP 约 9 帧，高斯泼溅超过 100 帧。实时，让高斯点从一篇论

文变成了一种可以在 VR 里走动的媒介。和第七讲一样的道理：速度改变了事物的本质。

Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

38. Under the hood: how splats are trained

Training. You start from the rough 3D points that photo-matching software, structure from motion, finds in your photos. Then repeat: render, compare with the photos, nudge every blob's position, shape and colour. Every so often, split blobs that are too big, clone blobs where detail is missing, and delete ones that don't contribute. On one GPU, a room takes roughly twenty to forty minutes. Generated worlds like Marble and Atlas skip the photos: a model imagines the views first, then lifts them to splats.

训练：从运动恢复结构（SfM）在照片里找到的粗略 3D 点开始。然后反复：渲染、与照片比较、微调每个高斯的位置、形状、颜色。每隔一段时间，把太大的高斯分裂、在缺细节的地方复制、删除没用的。一个房间在一块 GPU 上大约二三十分钟到四十分钟。Marble 和 Atlas 这样的生成世界跳过拍照：模型先想象出各个视角，再提升成高斯点。

Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

39. Capture or generate: two roads to the same file

So there are two roads to the same file. Capture: take sixty to a hundred and fifty photos or a video, match them to find camera positions, optimise splats for half an hour. Faithful, but you need to be there. Generate: give text, one image or a few photos; a model imagines the missing views and lifts them to 3D. You need nothing but a prompt, but it invents what it can't see. Atlas blurs the line: the more photos you give it, the less it invents.

所以到同一个文件有两条路。拍摄：拍 60 到 150 张照片或一段视频，匹配求出相机位置，优化半小时。忠实，但你必须到现场。生成：给文字、一张图或几张照片，模型想象缺失视角再提升为 3D。只需提示词，但看不到的部分是它编的。Atlas 模糊了界限：照片越多，编的越少。

40. This autumn, splats became a normal file format

Here's why I think 3D is really back. This autumn, splats stopped being a research format. The Blender 5.3 alpha imports and renders splats natively, in Eevee and Cycles. three.js, which many of you use for browser games, now has a built-in splat renderer. Khronos, the group behind glTF, has a splat extension.

PlayCanvas shipped SuperSplat 3.0 for editing. Companies are streaming 4D splats of live sport. For your projects, that means a splat world can go straight into your normal tools.

我为什么觉得 3D 真的回来了：今年秋天高斯点不再是研究格式。Blender 5.3 测试版原生导入并在 EEVEE 和 Cycles 里渲染高斯点。很多同学做浏览器游戏用的 three.js 内置了高斯点渲染器。制定 glTF 的 Khronos 有了高斯点扩展。PlayCanvas 发布了编辑器 SuperSplat 3.0。还有公司在直播体育赛事的 4D 高斯点。对你们的项目来说，高斯点世界可以直接进入常用工具。

41. Useful: phone – splat – Blender, in one evening

If you want to try this tonight. Capture: walk slowly around the subject, sixty to a hundred and fifty photos or a one-minute video, even light, no people walking through. Train with a desktop tool like Postshot or Brush, nerfstudio's gsplat if you like code, or just a phone app like Polycam or Scaniverse. Clean it up in SuperSplat in the browser: delete the floating blobs and the sky. Then use it in Blender 5.3 or the browser. One tip: a splat keeps the light you captured, so shoot at the time of day you want the scene to feel like.

今晚就可以试：拍摄时绕着物体慢慢走，60 到 150 张照片或一分钟视频，光线均匀，不要有人走过。用 Postshot、Brush 等桌面工具训练，喜欢代码可以用 nerfstudio 的 gsplat，或者直接用 Polycam、Scaniverse 等手机 App。在浏览器的 SuperSplat 里清理漂浮物和天空。然后放进 Blender 5.3 或浏览器。一个提示：高斯点保留了拍摄时的光，所以在你想要的时段拍。

42. Three art prompts for a splat

Three art prompts. First, a place about to disappear: a shop, your grandparents' flat, a street before redevelopment. A splat freezes a place the way a photo freezes a moment. Second, an impossible join: give Atlas or Marble two unrelated photos and let it invent what connects them, and look at what the model thinks belongs in between. Third, a level for your Assignment 3 game: scan a real corner of campus, then break it on purpose.

三个创作提示。一、即将消失的地方：一家店、祖父母的家、改造前的街道。高斯点冻结一个地方，就像照片冻结一个瞬间。二、不可能的连接：给 Atlas 或 Marble 两张无关的照片，让它想象连接它们的走廊，看看模型认为中间该有什么。三、作业三游戏的关卡：扫描校园真实的一角，然后故意打破它。

43. A photo freezes a moment. A splat freezes a place. What should you be allowed to scan, and who owns the copy?

Three minutes. A photo freezes a moment; a splat freezes a place, walkable, from any angle. What

should you be allowed to scan? Your room, fine. A shop? A temple? A museum gallery full of other people's art? And what about a stranger's home, reconstructed from photos they posted online, which Atlas says needs only two or three? Who owns the copy?

讨论三分钟。照片冻结瞬间，高斯点冻结一个可以从任何角度走进的地方。你应该被允许扫描什么？自己的房间没问题。商店？寺庙？满是别人作品的美术馆？一个陌生人的家呢——用他发在网上的照片重建，Atlas 说两三张就够。复制品属于谁？

44. Robots that learn by watching

Story three. In Lecture 07 a robot learned one dance from one phone video. Tonight: robots learning many everyday tasks from watching people, and from simulators built out of video.

第三个故事。第七讲里机器人从一段手机视频学会了一支舞。今晚：机器人通过观看人类、以及从视频构建的模拟器，学会许多日常任务。

45. Where we left robots: Lecture 07 and Lecture 03

Where we left robots. In Lecture 07, the K-pop humanoid: estimate the 3D pose from a phone video, retarget it onto the robot's body, and train a controller in simulation. In Lecture 03: an image location and a robot movement use different coordinate systems. Copying a dance is retargeting one motion. Tonight is harder: picking things up, opening fridges, wiping tables, from video of people.

我们之前讲到的机器人：第七讲的 K-pop 人形机器人，从手机视频估计 3D 姿态，重定向到机器人身体，在模拟中训练控制器。第三讲：图像位置和机器人动作用的是不同坐标系。复制一支舞只是重定向一个动作。今晚更难：从人的视频学拿东西、开冰箱、擦桌子。

46. Under the hood: what is inside a VLA?

Under the hood, what is a VLA? Chelsea Finn's Stanford course shows the common recipe. Start from a vision-language model, which already knows objects and words from the internet. Attach an action expert, a small network that uses diffusion or flow matching, the same generation method from Lecture 03, to output continuous joint commands instead of pixels. Train on robot demonstrations, plus web tasks so it doesn't forget language, plus human video. It outputs a short chunk of actions at a time.

VLA 内部是什么？斯坦福 Chelsea Finn 的课程展示了通用做法：从视觉语言模型开始，它已经从互联网学会了物体和词语。接上一个“动作专家”小网络，用扩散或流匹配（第三讲的生成方法）输出连续的关节指令而不是像素。训练数据包括机器人演示、网络任务（防止忘掉语言）和人类视频。每次输出一小段动作序列。

Adapted from Stanford CS 224R Deep RL, Lecture 17 "RL for Robot Foundation Models" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

47. Under the hood: why predict a chunk of actions?

Why chunks? A robot arm running at fifty hertz has to decide every twenty milliseconds. If it decides one tiny step at a time, it hesitates and jitters, especially when the demonstrations disagree. So modern policies predict about a second of actions at once, execute part of it, then re-plan. It's the same idea as generating a whole clip rather than one frame: you get smoother, more coherent motion.

为什么一次输出一段？50 赫兹的机械臂每 20 毫秒就要决策一次。一步一步决定会犹豫、抖动，尤其是演示数据彼此不一致时。所以现代策略一次预测约一秒的动作，执行一部分再重新规划。和生成整段视频而不是单帧是同一个道理：动作更平滑、更连贯。

Adapted from Stanford CS 224R Deep RL, Lecture 2 "Imitation Learning" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

48. Under the hood: small mistakes compound

The classic problem with learning by copying. The robot imitates the expert, but never perfectly. One small error takes it to a situation the expert never demonstrated, so it has no idea what to do there, and makes a bigger error. Errors compound. It's exactly the shape of drift in long video from Lecture 07: train on clean data, then at test time you feed on your own mistakes.

模仿学习的经典问题：机器人模仿专家，但永远不完美。一个小错误把它带到专家从未演示过的状态，它不知道该怎么办，于是犯更大的错误，误差累积。这和第七讲长视频的漂移完全同形：用干净数据训练，测试时却要自己犯的错。

Adapted from Stanford CS 224R Deep RL, Lecture 2 "Imitation Learning" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

49. The data problem: can robots learn from videos of people?

And the big one: data. Language models had the whole internet. Robots don't. Teleoperation, a person steering the robot, gives precise data but slowly and expensively. Videos of people are everywhere, but a human hand isn't a robot gripper, and our bodies move differently. Finn's slide calls it the embodiment gap. Both of tonight's robot stories are about bridging it.

最大的问题：数据。语言模型有整个互联网，机器人没有。遥操作（人操控机器人）数据精确，但慢且贵。人

类视频到处都是，但人手不是机械爪，身体运动方式也不同。Finn 的幻灯片称之为“具身差距”。今晚两个机器人故事都是关于如何跨越它。

*Adapted from Stanford CS 224R Deep RL, Lecture 2 "Imitation Learning" (Chelsea Finn, Spring 2026).
cs224r.stanford.edu*

50. $\lambda 0$ (26 Oct 2026): a humanoid that learned from people

$\lambda 0$, released yesterday by Alibaba's mapping company Amap. It's a VLA for a humanoid's whole body: walking, bending, carrying, using both hands. The robot is a Unitree G1 with five-fingered hands. What's new is the data: most of what it learned came from a dataset called HumanVerse, five hundred hours of people doing everyday things, with no robot present.

$\lambda 0$ ，阿里巴巴旗下高德昨天发布。这是控制人形机器人全身的 VLA：走路、弯腰、搬东西、双手操作。机器人是带五指手的宇树 G1。新意在数据：它学到的大部分来自 HumanVerse 数据集，500 小时人类日常活动，录制时没有机器人在场。

51. HumanVerse: wear the robot's eyes

How do you make human video useful for a robot? You wear the robot's eyes. In HumanVerse, people wear a GoPro on the neck at the same height as the robot's head camera, a body tracker recording twenty-four joints, and hand tracking with twenty-one joints per hand. Then they just live: tidy, carry boxes, water plants. The team estimates it's about 3.4 times faster than collecting the same data by teleoperating a robot.

怎样让人类视频对机器人有用？戴上机器人的眼睛。在 HumanVerse 里，人们把 GoPro 挂在脖子上，高度和机器人头部摄像头一致，身体追踪 24 个关节，手部每只 21 个关节。然后正常生活：收拾、搬箱子、浇花。团队估计比遥操作采集快约 3.4 倍。

52. $\lambda 0$: three stages, from human hands to a robot body

Training happens in three stages. One: learn how hands interact with objects from seven existing first-person human datasets. Two: the key step, translate HumanVerse into the robot's point of view and the robot's joints, so a person's reach becomes a robot's reach. Three: finish with a small amount of real robot data, six hundred demos, a hundred and fifty per task. At run time, the robot gets its camera image, an instruction and its body sense, and outputs whole-body motion.

训练分三阶段。一：从七个第一人称人类数据集学习手与物体的交互。二：关键一步，把 HumanVerse 转换到

机器人的视角和关节，让人的伸手变成机器人的伸手。三：用少量真实机器人数据收尾，600 条演示，每个任务 150 条。运行时机器人输入摄像头画面、指令和本体感觉，输出全身动作。

53. $\lambda 0$, by the numbers

The numbers. On four real tasks, the strongest baseline, Physical Intelligence's $\pi 0.5$, succeeded 40 percent of the time. $\lambda 0$: 62.5 percent, twenty-five out of forty. The most interesting part is the ablation. Remove stage two, the human-to-robot translation, and it falls to 22.5. Remove all human data, 15 percent. So the human video isn't decoration; it's most of the skill. Also note that 62.5 is far from good enough for a home. Keep that in mind.

数据：四个真实任务上，最强基线 Physical Intelligence 的 $\pi 0.5$ 成功率 40%。 $\lambda 0$ 是 62.5%，40 次成功 25 次。最有意思的是消融实验：去掉第二阶段（人到机器人的转换）降到 22.5%；去掉所有人类数据只有 15%。人类视频不是装饰，而是大部分技能的来源。同时注意，62.5%离家用还差得远。

54. Gemini 4 Argon (13 Oct): watch a video, write a simulator

A completely different route, from Google two weeks ago. Gemini 4 Argon isn't a robot policy; it's a general model that's very good at physical reasoning and code. You give it a robot camera feed, a video of a person, or even a hand-drawn sketch, and it writes a MuJoCo physics simulation of that scene. Then it writes control code, practises in the simulator and transfers the result. In one of their demos, simulation-generated data raised real-world success from 56 to 75 percent.

完全不同的路线，谷歌两周前发布。Gemini 4 Argon 不是机器人策略，而是物理推理和编程很强的通用模型。给它机器人摄像头画面、人的视频，甚至一张手绘草图，它就写出这个场景的 MuJoCo 物理模拟。然后写控制代码，在模拟器里练习，再迁移到真实机器人。一个演示里，模拟生成的数据把真实成功率从 56%提高到 75%。

55. Argon demos: video – simulation – robot

Four of their demos. [Click through.] A sketch of a humanoid throwing a ball becomes a simulation. A person's first-person video becomes a simulated twin of their kitchen. Argon steers a video world model, Gemini Omni, to spell its own name with Scrabble tiles, which is Story 1 and Story 3 meeting. And a cheap SO-101 arm, the kind you can buy for a few hundred dollars, gets faster at a pick-and-place task by practising: 59 seconds, then 23, then 17.

四个演示。草图变成人形机器人投球的模拟；第一人称视频变成厨房的模拟孪生；Argon 操控视频世界模型 Gemini Omni 用拼字游戏的字母块拼出自己的名字，这是故事一和故事三相遇；一个几百美元就能买到的 SO-

101 机械臂通过练习越来越快：59 秒、23 秒、17 秒。

56. Watch one person play cornhole, then do it

And last week, the demo that made the headlines. Argon watched an ordinary video of a person playing cornhole, the American bean-bag throwing game. It mapped the space, worked out the whole-body joint motion, wrote the simulation itself, trained the skill in simulation, and transferred it to Appttronik's Apollo humanoid. In a real trial, Apollo landed seven throws in a row. Google says this 'video-to-sim-to-real' pipeline turns weeks of manual simulation work into hours.

上周上头条的演示：Argon 看了一段普通的人玩沙包投掷游戏的视频，绘制空间、算出全身关节运动、自己写出模拟、在模拟中训练，再迁移到 Appttronik 的 Apollo 人形机器人。真实测试中连续七次命中。谷歌说这条“视频→模拟→现实”的流程把几周的手工模拟工作缩短到几小时。

57. Gemini Robotics 2 (July): one model for the whole body

For context, the model we met in Lecture 03, now in motion. Gemini Robotics 2, from July, controls a humanoid's whole body, not just the arms over a table. An instruction like 'put the watering can in the green bin on the bottom shelf' means walking, picking, walking, placing. A separate reasoning model, ER 2, plans the steps and checks progress, and the VLA moves. It can adapt to a new robot body with typically fewer than two hundred examples.

背景：第三讲见过的模型，现在动起来了。七月的 Gemini Robotics 2 控制人形机器人全身，而不只是桌面上的手臂。“把浇水壶放进底层的绿色箱子”意味着走、拿、走、放。另一个推理模型 ER 2 负责规划和检查进度，VLA 负责动作。适应一个新机器人身体通常只需不到 200 个示例。

58. π 0.7 and the Pi API: robot brains as a service

Physical Intelligence, the startup behind the π models. π 0.7 in April could be steered with more than words: you can say faster, or higher quality, or show a subgoal image. It folded laundry on a robot that had no laundry training data. And on the 13th of October they opened a beta Pi API: upload your robot's demonstrations, they fine-tune and host the model. Robot brains become a cloud service. Their older π 0 and π 0.5 weights are open, and that's what λ 0 was compared against.

Physical Intelligence， π 系列模型背后的创业公司。四月的 π 0.7 可以用文字以外的方式控制：更快、更高质量，或给一张子目标图片。它在没有叠衣服训练数据的机器人上叠好了衣服。10 月 13 日他们开放了 Pi API 测试：上传你的机器人演示数据，他们微调并托管模型。机器人脑变成云服务。旧版 π 0 和 π 0.5 权重是开源的， λ 0 就是和它比较的。

59. Under the hood: why copying plateaus around 80%

Under the hood, why isn't λ_0 's 62 percent, or anyone's 80, good enough? Finn's lecture makes the point clearly: VLAs trained only by imitation tend to plateau around 80 percent. A robot in a home needs 99 percent or more, because a robot that drops one cup in five is not a product. And copying can't make you better than your teachers. So the frontier is letting robots practise and learn from their own attempts, which is reinforcement learning.

为什么 λ_0 的 62% 甚至 80% 都不够？Finn 的课讲得很清楚：只靠模仿训练的 VLA 往往停在 80% 左右。家里的机器人需要 99% 以上，五次摔一个杯子的机器人不是产品。而且模仿不可能超过老师。所以前沿是让机器人练习、从自己的尝试中学习，也就是强化学习。

Adapted from Stanford CS 224R Deep RL, Lecture 17 "RL for Robot Foundation Models" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

60. Practice beats copying: RL roughly doubles throughput

Here is the evidence from the same lecture. Physical Intelligence's π -star-0.6 learned a value function, roughly a sense of 'is this going well?', and fine-tuned on what worked. Throughput roughly doubled compared with imitation alone, on real tasks like making espresso. And this is where world models come back. Practising on real dishes is slow and expensive. Practising inside a world model or a simulator built from video, like Argon's, is cheap.

同一讲的证据：Physical Intelligence 的 $\pi^*0.6$ 学习了价值函数——大致是“现在进展顺利吗”的感觉——然后在成功的经验上微调。在做浓缩咖啡等真实任务上，吞吐量比只用模仿提高约一倍。世界模型在这里回归：在真实的盘子上练习又慢又贵，在世界模型或从视频构建的模拟器（如 Argon）里练习很便宜。

Adapted from Stanford CS 224R Deep RL, Lecture 17 "RL for Robot Foundation Models" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

61. Three ways to teach a robot, 2026

So in 2026 there are three ways to teach a robot. Steer it: teleoperation, precise but slow, and Physical Intelligence's open models were pretrained on more than ten thousand hours of it. Show it: human video, cheap and everywhere, but a different body, as with λ_0 . Imagine it: simulation and world models, unlimited practice, but the sim-to-real gap from Lecture 07. Every serious system now mixes all three.

2026 年教机器人的三种方式。操控它：遥操作，精确但慢，Physical Intelligence 的开源模型用了一万多小时这样的数据预训练。示范给它看：人类视频，便宜、到处都有，但身体不同，比如 λ_0 。让它想象：模拟和世界

模型，无限练习，但有第七讲讲过的模拟到现实差距。现在每个严肃的系统都三者结合。

62. Be the VLA

A game. One volunteer per pair is the robot: eyes closed, moves only when told. The partner gives one-sentence commands, like 'move your right hand twenty centimetres to the left', and the goal is to put a pen into a cup. Count the commands. Then let the robot open its eyes and try again with one instruction: 'put the pen in the cup'. What changed? Vision plus language is what lets one sentence replace twenty low-level commands. That's the V and the L in VLA.

游戏：每两人一组，一人当机器人，闭眼，只按指令动。另一人用一句话发指令，比如“右手向左移动 20 厘米”，目标是把笔放进杯子。数一数用了几条指令。然后让“机器人”睁眼，只给一条指令：“把笔放进杯子”。有什么不同？视觉加语言让一句话代替二十条底层指令。这就是 VLA 里的 V 和 L。

63. If robots learn from first-person video of people's daily lives, whose bodies and homes become the training data?

Three minutes. If robots learn from first-person video of people's daily lives, whose bodies and whose homes become training data? HumanVerse paid people to record. But there are millions of hours of first-person video online already. Who should be paid or credited? What should never be recorded? And personally, would you wear the camera for a week? For how much?

讨论三分钟：如果机器人从人类日常生活的第一人称视频学习，谁的身体和家会变成训练数据？HumanVerse 是付钱请人录的，但网上已经有数百万小时第一人称视频。谁应该得到报酬或署名？什么永远不该被录下？你个人愿意戴一周摄像头吗？多少钱？

64. Does a world model need to draw pixels?

Story four, a real disagreement. Everything so far draws pixels or splats. Some people think that's the wrong goal entirely.

第四个故事，一个真实的分歧。目前为止所有东西都在画像素或高斯点。有些人认为这个目标本身就错了。

65. Two camps

Two camps. One generates the world: predict the next frame in pixels or splats. The advantage is you can watch the prediction, train robots inside it, and also sell it as video. Genie, Omni, Cosmos, Atlas. The other predicts in latent space: predict an abstract summary of what happens next, never the pixels. Why

waste effort drawing every leaf and reflection, if what matters for acting is where the cup goes? That's Yann LeCun's position, with JEPA.

两个阵营。一派生成世界：用像素或高斯点预测下一帧。好处是你能看到预测、在里面训练机器人，还能当视频卖。Genie、Omni、Cosmos、Atlas。另一派在隐空间预测：预测下一刻的抽象概括，从不画像素。如果行动只需要知道杯子往哪去，何必费力画每片叶子和倒影？这是杨立昆的 JEPA 立场。

66. NVIDIA Cosmos 3: one open model to reason, simulate and act

The strongest open version of the first camp is NVIDIA's Cosmos 3, from the end of May. It's one model with two towers: a reasoner, which is a vision-language model, and a generator that outputs video and robot actions. There's a 32-billion-parameter version for data centres and an 8-billion one for a workstation. NVIDIA released the weights, training code and datasets under a licence that allows commercial use. What is it for? Mostly making synthetic training data and testing robot policies before they touch real hardware.

第一阵营里最强的开源模型是 NVIDIA 五月底发布的 Cosmos 3。一个模型两个塔：推理塔是视觉语言模型，生成塔输出视频和机器人动作。有 320 亿参数的数据中心版和 80 亿参数的工作站版。NVIDIA 开放了权重、训练代码和数据集，许可证允许商用。用途主要是生成合成训练数据，以及在接触真实硬件前测试机器人策略。

67. Why generate pixels? Because cameras are what robots see

The argument for pixels. A robot's camera sees pixels and depth, so to train or test it in imagination, someone must produce pixels and depth. Atlas does exactly that: from twenty-four frames of phone video per room, it generates what any robot would see along any path. Waymo uses the same idea for crashes it can't film. And pixels are checkable: you can watch what the model believes and catch it when it's wrong.

支持像素的论点：机器人的摄像头看到的是像素和深度，所以要在想象中训练或测试它，就得有人生成像素和深度。Atlas 正是这样：每个房间 24 帧手机视频，就能生成机器人沿任意路径看到的画面。Waymo 用同样思路处理无法拍摄的事故。而且像素是可检查的：你能看到模型“相信”什么，发现它错的时候。

68. Why not pixels? JEPA predicts the meaning, not the picture

The argument against. In JEPA, you hide part of a video and predict the representation of the missing part, an abstract summary, not its pixels. Meta's V-JEPA 2 was pretrained on more than a million hours of video, then adapted with under 62 hours of unlabelled robot video into a model that can plan robot actions. LeCun left Meta and founded AMI Labs in Paris, which raised a 1.03-billion-dollar seed round in

March to build this. His argument: a model that must draw every pixel wastes its capacity on details that don't matter for acting.

反方论点：JEPA 里把视频遮住一部分，预测被遮部分的表示——抽象概括，而不是像素。Meta 的 V-JEPA 2 用超过一百万小时视频预训练，再用不到 62 小时无标注机器人视频适配成能规划机器人动作的模型。杨立昆离开 Meta，在巴黎创立 AMI Labs，三月融资 10.3 亿美元种子轮来做这件事。他的论点：必须画出每个像素的模型，把能力浪费在与行动无关的细节上。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423 Facts: Meta V-JEPA 2 (2025); Toronto Starts on AMI Labs, Sept 2026.

69. V-JEPA 2-AC: plan a robot move without drawing the future

Here is what that looks like in practice, from Meta's V-JEPA 2 paper. You show the robot a goal image: the cup should end up here. The model imagines many possible action sequences, but only in embedding space; it never draws a single frame. It picks the sequence whose predicted embedding is closest to the goal's, executes one step, looks again and re-plans. That's model-predictive control, with a world model that never makes a picture.

实际效果，来自 Meta 的 V-JEPA 2 论文。给机器人一张目标图：杯子应该在这里。模型在嵌入空间里想象许多动作序列，从不画出一帧画面；选出预测嵌入最接近目标的序列，执行一步，再看、再规划。这就是模型预测控制，用的是一个从不生成图像的世界模型。

70. Under the hood: the two losses, side by side

CMU's slide puts the two losses side by side. The generative loss compares the predicted next observation with the real one, in pixel space. The JEPA loss compares the predicted embedding with the real embedding, in a learned space. The structure is the same, state plus action gives next state. The whole fight is about what you compare: pictures or meanings.

CMU 这页把两种损失并排放。生成式损失在像素空间比较预测的下一观测和真实的；JEPA 损失在学到的空间比较预测的嵌入和真实嵌入。结构一样，状态加动作得到下一状态。整个争论在于比较什么：画面还是含义。

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

71. How would you know a world model really “understands”?

How would you know a world model really understands the world? Four tests. The round trip from Lecture 03. A counterfactual: push the cup instead of lifting it, does the outcome change in the right way? Long horizon: still consistent after ten minutes, an hour? And real action: does a robot that planned inside the model succeed in the real room? Only the last test works for both camps, which is why robotics has become the referee of this debate.

怎么知道世界模型真的“理解”世界？四个测试。第三讲的往返测试。反事实：推杯子而不是拿起杯子，结果会正确地改变吗？长时程：十分钟、一小时后还一致吗？真实动作：在模型里规划的机器人在真实房间里能成功吗？只有最后一个测试对两个阵营都适用，所以机器人成了这场争论的裁判。

72. A video generator is a world model.

Debate. The motion: a video generator is a world model. Left side for, right side against. Two speakers per side, one minute each. Use tonight's examples: Genie, Omni, Atlas, Cosmos, JEPA, the robot results. Then we vote.

辩论。辩题：“视频生成器就是世界模型。”左边正方，右边反方，每边两位，每人一分钟。用今晚的例子：Genie、Omni、Atlas、Cosmos、JEPA 和机器人结果。然后全场举手投票。

73. Arguments to start from

Some arguments to start from, if your side is stuck. For: Genie 3 already takes actions and keeps memory; Waymo uses it for real cars; Argon steers Omni to move objects, which means the model predicts consequences; and to draw the future correctly at all, you need some model of the world. Against: most video generators take no actions, they predict but don't respond; they forget and drift, as we saw in Lecture 07; a beautiful future can be wrong; and LeCun's point, the world is not its pixels.

如果卡住了，可以从这些论点开始。正方：Genie 3 已经接受动作并保有记忆；Waymo 用它训练真车；Argon 操控 Omni 移动物体，说明模型能预测后果；要正确地画出未来，本身就需要某种世界模型。反方：大多数视频生成器不接受动作，只预测不回应；会遗忘、漂移、违反物理（第七讲）；漂亮的未来可能是错的；杨立昆的观点：世界不等于它的像素。

74. Four stories, one line each

Four stories, one line each. Add an action to a video model and it becomes a place, and the hard part is memory. Keep the geometry and the world becomes a file you can open, edit and sell. Robots now learn mostly from people and from imagined worlds, but still fail about one time in three. And whether a

world model must draw the world at all is still open; robots will decide.

四个故事，各一句话：给视频模型加上动作，它就变成一个地方，难点在记忆。保留几何，世界就变成可以打开、编辑、出售的文件。机器人现在主要从人类和想象的世界中学习，但仍然三次失败一次。世界模型是否必须画出世界仍未定论，机器人会给出答案。

75. Tonight – your assignments (both due Nov 30)

How tonight connects to your assignments, both due November 30th. For Assignment 3, your game: sketch a level with a world model or Marble, or scan a real place as a splat and load it into three.js or Unity. For Assignment 2, a splat or a world-model clip can be your set, just say clearly which tool made which part. And some final-project ideas: an explorable memory, a world you can only enter once, a robot that learns your gesture.

和作业的联系（两个作业都是 11 月 30 日截止）。作业三的游戏：用世界模型或 Marble 草拟关卡，或把真实地点扫描成高斯点放进 three.js 或 Unity。作业二：高斯点或世界模型片段可以当布景，清楚说明哪部分用了哪个工具。期末项目想法：可探索的记忆、只能进入一次的世界、学会你手势的机器人。

76. Three readings for this week

Three readings. One, World Labs' Atlas post: read the spatial-context part and the robotics part. Two, the $\lambda 0$ project page: focus on HumanVerse, the three training stages and the ablation. Three, a TechCrunch piece on Genie 3 and why DeepMind thinks world models matter for general AI; read it critically, with tonight's debate in mind. Optional for the technical students: Google's Gemini 4 Argon robotics post.

三篇阅读：一、World Labs 的 Atlas 文章，读空间上下文和机器人部分。二、 $\lambda 0$ 项目页，重点看 HumanVerse、三阶段训练和消融实验。三、TechCrunch 关于 Genie 3 的文章，看 DeepMind 为何认为世界模型对通用 AI 重要。技术方向同学可选读谷歌的 Gemini 4 Argon 机器人文章。

77. Quiz 09 · five short questions

Quiz 09, five short questions, same format as before: two technique, two art, one reading reflection. Due by the end of the semester, on the course site.

第九次测验，五个简短问题，格式不变：两道技术、两道艺术、一道阅读反思。学期末前在课程网站提交。

78. Every demo from tonight, in one place

Every demo and source from tonight is on this slide and on the course page, so you can go back to them.

今晚所有演示和来源都在这一页和课程网页上，方便回看。

79. Critique, evaluation, provenance, and responsible release

Next week we step back. Now that we can make worlds, voices and bodies, how do we judge them, how do we label them, and how do we decide what to release? Bring one piece you've made this term.

下周我们退一步：既然我们能造世界、声音和身体，该如何评价、如何标注、如何决定发布什么？请带一件你这学期做的作品。

80. Worlds, Splats, and Bodies

Worlds, splats and bodies. Thank you. Questions?

世界、高斯点和身体。谢谢大家，有问题吗？

81. Do it yourself

An appendix for your projects: tools, a workflow, and a review checklist.

附录：工具、一晚上的工作流程和检查清单。

82. Toolkit: what to use for each job

A toolkit, one row per job. For a world from a prompt: Marble online, or HY-World 2.0 if you have a big GPU. To try a playable world: Project Genie or Gemini Omni. To scan a place: Postshot, Brush, gsplat, or a phone app. Edit in SuperSplat. Use the result in Blender 5.3 or three.js. If you want to touch robots: openpi and LeRobot with an SO-101 arm, and MuJoCo for simulation. Check licences before commercial use.

工具表，每行一项工作。从提示词生成世界：在线用 Marble，或有大显卡用 HY-World 2.0。体验可玩世界：Project Genie 或 Gemini Omni。扫描场景：Postshot、Brush、gsplat 或手机 App。用 SuperSplat 编辑。在 Blender 5.3 或 three.js 里使用。想碰机器人：openpi 和 LeRobot 配 SO-101 机械臂，用 MuJoCo 模拟。商用前查许可证。

83. Workflow: a walk-in level for your game, in one evening

A one-evening workflow for a walk-in level. Pick a place. Scan it or generate it. Clean it in SuperSplat and set the scale and floor. Load it into three.js with Spark, or Blender. Add a simple collider so you don't

walk through walls, and one interaction. Then playtest: walk away, come back, try to break it. Roughly three hours in total. And as in Lecture 07, keep your source photos and export settings.

一晚上做一个可走进去的关卡：选地点；扫描或生成；在 SuperSplat 清理并设置比例和地面；载入 three.js (Spark) 或 Blender；加一个简单碰撞体和一个交互；然后测试：走开、回来、试着弄坏它。总共大约三小时。像第七讲说的，保留原始照片和导出设置。

84. Review checklist for a generated world

And a checklist before you submit a generated world. Consistency. Scale and physics. Which parts were invented by the model, and are they labelled? Rights: did you have permission to scan, and whose photos? Export details. And disclosure, which leads into next week.

提交生成世界前的检查清单：一致性；比例和物理；哪些部分是模型编的，有没有标注；权利：是否获准扫描、用了谁的照片；导出细节；披露——这正是下周的主题。