AMCC 5160 · Lecture 11 transcript

The Semester in One Evening: a review of Lectures 01–10 · 10 November 2026

1. The Semester in One Evening

Good evening, and welcome to Week 11. This is our last lecture before the final presentations, so tonight I won't add anything new. We'll look back instead, and in more detail than usual. For ten weeks we've moved fast: neurons, diffusion, editing, customization, quantization, minute-long video, sound, avatars, worlds, robots, and critique. Tonight we'll walk back through every lecture in order, with the technical content first: the architectures, the equations and the numbers. For each one I'll put the original slides back on screen, the exact slides you saw that night, and we'll ask what you should still remember and how it connects to your projects. On almost every slide you'll see the lecture and slide number, like "L03 · 54". On the website each of those is a link, so this deck works as an index to the whole semester. After the review and a break, we have the last weekly presentations.

欢迎来到第十一周。这是期末展示前的最后一讲，今晚不讲新内容，而是比平时更细致地回顾。十周里我们讲了神经网络、扩散、编辑、定制、量化、一分钟视频、声音、数字人、世界模型、机器人和评论。今晚按顺序回到每一讲，技术内容优先：架构、公式和数字。把当晚的原始幻灯片重新放到屏幕上，问你还应该记住什么、它和你的项目有什么关系。几乎每页都标着讲次和页码，比如“L03 · 54”，网站上都可以点开，所以这份幻灯片也是整个学期的索引。回顾和休息之后是最后一轮每周展示。

Revisit: L01 · 1 (How machines learn to make moving worlds), L03 · 1 (Inside the generator), L05 · 1 (Making the Model Yours), L07 · 1 (What's New in AI Motion), L08 · 1 (When AI Video Learned to Talk), L09 · 1 (Worlds You Can Walk Into, Robots That Learn by Watching)

2. Ten Tuesdays on one line

Here are ten Tuesdays on one line, one sentence each. Week 1: a generated video is a learned prediction about change. Week 2: write the edit contract, meaning what changes, what stays and what adapts. Week 3: the model predicts and the sampler takes a step; seed, guidance and steps are different knobs. Week 4: video has two kinds of time, and the artist's intention comes first. Week 5: whose model is it? Week 6, Yaofu's talk: store less, using fewer bits per number. Week 7: the unit of AI video became the scene, and tools can vanish. Week 8: sound and picture share one clock, and we asked who the performer is. Week 9: add an action and video becomes a place. Week 10: is it AI, is it good, is it fair, is it allowed? If you can say these ten sentences and give an example for each, you're in good shape.

十个星期二，一句话一个。第 1 周：生成视频是对变化的学习预测。第 2 周：写编辑契约，改什么、保留什么、什么随之调整。第 3 周：模型预测，采样器走一步；种子、引导、步数是不同的旋钮。第 4 周：视频有两种时间，艺术意图在先。第 5 周：模型是谁的？第 6 周（Yaofu）：用更少的比特存储。第 7 周：AI

视频的单位变成了场景，工具会消失。第 8 周：声音和画面共用一个时钟，表演者是谁？第 9 周：加入动作，视频就变成了地方。第 10 周：是 AI 吗？好吗？公平吗？允许吗？能说出这十句话并各举一例，就准备得不错了。

3. Neural networks to AI filmmaking

Lecture 1, September 1st. The title slide said: how machines learn to make moving worlds. It was the longest lecture of the term, a hundred and two slides, and it tried to give you a mental model that would survive the next product release. We went from a single neuron to CNNs, attention and transformers, latent models, diffusion and flow, video as a temporal system, the 2026 landscape with Seedance and MiniMax H3, and finally a full AI filmmaking workflow. The question that night was simple: what is a generated video, really? The answer: a learned prediction about change. Let's look at four ideas from that night that you've used every week since.

第 1 讲，9 月 1 日。标题是：机器如何学会制作运动的世界。这是本学期最长的一讲，102 页，目标是给你一个能撑过下一次产品发布的心智模型。从单个神经元到 CNN、注意力和 Transformer、潜变量模型、扩散和流、作为时间系统的视频、2026 年的格局（Seedance、MiniMax H3），最后是完整的 AI 电影制作流程。那晚的问题：生成视频到底是什么？答案：对变化的学习预测。下面看那晚四个你每周都在用的想法。

Revisit: L01 · 1 (How machines learn to make moving worlds), L01 · 7 (A generated video is a learned prediction about change), L01 · 102 (Technical literacy becomes creative agency)

4. Every generator in four stages

This is the map from Lecture 1, and it held all semester. Data supplies examples. Representation decides what becomes compact and comparable: latents and tokens. The model learns relationships in that representation. Sampling turns what was learned into one particular output. The useful part is the debugging question. When a face drifts, ask which stage it is: missing data, a weak identity representation, too little temporal modelling, or an inference setting? And remember Lecture 4: data is never neutral. Filtering decides what reaches training, so a model's defaults reflect choices other people made.

这是第 1 讲的地图，整个学期都适用。数据提供例子；表示决定什么被压缩、可比较（潜变量和 token）；模型在这个表示里学习关系；采样把学到的东西变成一个具体输出。最有用的是诊断问题：脸漂移了，是哪一步？缺数据、身份表示弱、时间建模不足，还是推理设置？也记住第 4 讲：数据从不中立，筛选决定了什么进入训练，所以模型的默认结果反映了别人的选择。

Revisit: L01 · 8 (Every generator can be read through four stages), L04 · 15 (Training data shapes the result), L04 · 16 (A closer look at data filtering)

5. From one neuron to a video model, in five steps

Here's Lecture 1's first half compressed into one line. A neuron computes a weighted sum, adds a bias and applies

a nonlinearity, and training adjusts those weights to reduce error across examples. For diffusion, the target is the noise that was added. A CNN scans one learned filter across the image, because nearby pixels usually belong together. Attention lets every token choose what matters, so the red scarf in frame twenty can attend to the scarf in frame two. Attention alone doesn't know order, which is why position encodings like RoPE add time, height and width. Then a VAE compresses first and the generator learns to create in that smaller space. Diffusion and flow learn the path from noise to data. Every system you met afterwards, including Helios, LTX-2 and Genie, is a rearrangement of these five parts.

第 1 讲前半部分压缩成一行。神经元：加权求和、加偏置、过非线性，训练调整权重以减少误差，扩散模型的目标就是加进去的噪声。CNN：同一个学到的滤波器扫过整张图，因为相邻像素往往属于同一物体。注意力：每个 token 自己选择重要的东西，第 20 帧的红围巾可以关注第 2 帧的围巾；注意力本身不知道顺序，所以需要 RoPE 等位置编码加入时间、高、宽。VAE 先压缩，再在小空间里学习生成。扩散和流学习从噪声到数据的路径。之后遇到的每个系统，Helios、LTX-2、Genie，都是这五个部件的重新组合。

Revisit: L01 · 11 (Learning means reducing error across examples), L01 · 16 (A convolution scans one learned filter across the image), L01 · 26 (Queries ask; keys match; values contribute), L01 · 33 (An encoder and decoder negotiate a bottleneck), L01 · 36 (Training creates many controlled corruption levels)

6. Compress first, then create

Why do all these models work in a compressed space? Count. A ten-second clip at twenty-four frames per second is two hundred and forty frames before any sound. Attention compares every token with every other token, so cost grows with the square of length. Lecture 7's rule of thumb: six times the length means roughly thirty-six times the comparisons. So a video VAE squeezes space and time first. MiniMax H3 reported sixteen times spatial and four times temporal compression, and then groups the latent into patch tokens, small blocks of space-time. Bigger patches are cheaper but lose fine motion and text. That trade-off decided which details survived in your Assignment 1 films.

为什么这些模型都在压缩空间里工作？算一算：十秒、每秒 24 帧，就是 240 帧，还没算声音。注意力让每个 token 和所有 token 比较，成本随长度平方增长；第 7 讲的经验法则：长度六倍，比较次数约三十六倍。所以视频 VAE 先压缩空间和时间，MiniMax H3 报告空间 16 倍、时间 4 倍压缩，再切成时空小块 token。块越大越便宜，但会丢细微运动和文字。这个取舍决定了你们作业一里哪些细节留了下来。

Revisit: L01 · 43 (Time adds another expensive axis), L01 · 44 (Compress space and time before the transformer), L07 · 9 (Why five seconds was the limit)

7. Video tokens: patches, positions, and which tokens may talk

Here's the technical core of Lecture 1's video section. After the VAE, patchification groups latent values into tokens, and each token is a small block of space-time that carries content plus coordinates. Larger patches cut cost but lose fine motion and text, and smaller patches keep detail but lengthen the sequence. Attention

compares a set of vectors, so on its own it doesn't know left from right or before from after. Position encodings inject coordinates, and RoPE rotates query and key components so relative position changes similarity. In video that needs three axes, time, height and width, which is why H3 and HunyuanVideo describe spatiotemporal RoPE. Last, which tokens may talk to which: full spatiotemporal attention is the most flexible and grows quadratically, while sparse, sliding, factorized or block attention limits who attends to whom. HunyuanVideo 1.5 uses selective and sliding-tile attention. The question to ask of any architecture: what is a token, how does it know where it is, and who can it see?

第 1 讲视频部分的技术核心。VAE 之后，分块把潜变量组成 token，每个 token 是带着内容和坐标的一小块时空。块越大成本越低，但会丢失细微运动和文字；块越小保留细节，但序列更长。注意力比较的是一组向量，本身不分左右先后，位置编码注入坐标，RoPE 旋转 query 和 key 的分量，让相对位置影响相似度。视频需要时间、高、宽三个轴，所以 H3 和 HunyuanVideo 都描述了时空 RoPE。最后，哪些 token 能互相看：完整时空注意力最灵活但二次增长；稀疏、滑动、分解或分块注意力限制谁关注谁，HunyuanVideo 1.5 用了选择性和滑动分块注意力。问任何架构三个问题：token 是什么？它怎么知道自己在哪？它能看见谁？

Revisit: L01 · 45 (A token can represent a small block of spacetime), L01 · 29 (Attention alone does not know order), L01 · 46 (Full attention is powerful; sparse attention is practical)

8. A film is a sequence of decisions, not one generation

This is the second half of Lecture 1, and it was the workflow behind Assignment 1. Current models are shot engines, reference engines and editing engines. A film still needs dramatic structure, coverage, continuity, selection, sound and rhythm. So we used five gates: the brief, the world bible, the animatic, shot generation, and post-production. Each gate produces something you can inspect before you move on. Two rules from that night are worth repeating. Define change before style: "a courier must deliver one memory before the city deletes her identity" gives you action and consequence, while "cinematic cyberpunk woman" mostly gives you decoration. And use the least generative task that solves the shot: text-to-video when you want invention, image-to-video to keep a designed frame, first-last-frame to control endpoints, and video-to-video to keep blocking and timing.

这是第 1 讲后半部分，也是作业一背后的流程。现在的模型是镜头引擎、参考引擎和编辑引擎，但一部电影仍需要戏剧结构、覆盖镜头、连贯、选择、声音和节奏。所以有五道关：简报、世界设定、动态分镜、镜头生成、后期，每道关都产出可检查的东西。两条规则值得重复：先定义变化再谈风格，“信使必须在城市删除她的身份之前送出一段记忆”有动作有后果，“电影感赛博朋克女人”只是装饰；用能解决问题的最少生成的任务：要发明就用文生视频，要保留设计好的画面就用图生视频，控制首尾用首尾帧，保留走位和节奏用视频生视频。

Revisit: L01 · 85 (Five gates prevent expensive confusion), L01 · 86 (Define change before style), L01 · 90 (Use the least generative task that solves the shot)

9. Hold the shot constant; vary one decision

And this is the habit Lecture 1 ended on, which I've repeated almost every week since. Start from a known-good workflow and run it once before you add anything. Then hold the shot constant and vary one decision: camera strength, endpoint frame, guidance or motion reference. Generate three to six candidates and write your observations before you choose a favourite. Changing model, seed, prompt, duration and reference all at once creates variety, but almost no knowledge. Then keep the usable interval, not the prettiest thumbnail: watch at full speed, mark in and out points, and keep one representative failure. Finally, edit first and repair only what survives the cut. The last slide of that night said that technical literacy becomes creative agency. This slide is what that means in practice. The check-yourself question is the exercise from that night: design one eight-second shot before you open a generator.

这是第 1 讲结尾的习惯，此后几乎每周都在重复。从一个能跑通的工作流开始，先运行一次再加东西。然后固定镜头，只改一个决定：镜头运动强度、结尾帧、引导或运动参考。生成三到六个候选，先写观察再挑最喜欢的。同时改模型、种子、提示词、时长和参考，会带来变化，却几乎带不来知识。保留可用的区间，而不是最漂亮的缩略图：全速看、标出入点、保留一个代表性的失败。先剪辑，只修复剪辑后留下的部分。那晚最后一页说：技术素养变成创作能力。这一页就是它的实际含义。自测题就是那晚的练习：打开生成器之前，设计一个八秒镜头。

Revisit: L01 · 96 (Hold the shot constant; vary one decision), L01 · 95 (Change one variable after a known-good workflow runs), L01 · 97 (Keep the usable interval, not the prettiest thumbnail), L01 · 101 (Design one eight-second shot before opening a generator)

10. Generative editing

Lecture 2, September 8th: generative editing. We spent the whole evening with one object, an amber ceramic pear with a blue stem, and asked it to survive three things: a change of environment, an art direction for a fictional exhibition called AFTER RAIN, and a moving camera. The question was: when you ask for a change, what exactly should stay the same? It sounds like a technical question, but it turned out to be an artistic one, because "the same" has more than one meaning. This was also the night we first met flow matching, classifier-free guidance, ControlNet and LoRA, and the first metric that rewarded the wrong thing.

第 2 讲，9 月 8 日：生成式编辑。一整晚只围绕一个物体：琥珀色陶瓷梨和蓝色的梗。让它经受三件事：环境变化、为虚构展览“AFTER RAIN”做艺术指导、移动镜头。问题是：要求改变时，到底什么应该保持不变？听起来是技术问题，其实是艺术问题，因为“相同”不止一种意思。这一晚我们也第一次遇到流匹配、无分类器引导、ControlNet、LoRA，以及第一个奖励错误东西的指标。

Revisit: L02 · 1 (Generative Editing), L02 · 9 (The edit contract), L02 · 98 (The moment the pear returns)

11. The edit contract: change, keep, adapt

Remember the pear. The gallery becomes a forest. The edit contract had three columns. Change: the environment. Keep: the pear's identity, its pedestal, the framing. Adapt: the light and the reflections, because a bright gallery window reflected in a dark forest would be faithful and wrong. Then Lecture 2's subtle point: "keep it the same" has two meanings. Semantic sameness means a visitor recognizes the same sculpture. Exact sameness means identical pixels in a region, and only copying or compositing can promise that. Write this contract before you generate, and check each column separately afterwards. It works for a still image, a video shot, and a game asset in Assignment 3.

还记得那只梨吗？展厅变成森林。编辑契约有三栏。改变：环境。保留：梨的身份、底座、构图。调整：光线和反射，因为在黑暗森林里保留明亮展厅窗户的倒影，忠实却是错的。第 2 讲还有一个细节：“保持不变”有两种意思。语义上相同，观众认出是同一件雕塑；精确相同，某区域像素完全一样，只有复制或合成才能保证。生成前写好契约，生成后逐栏检查。静态图、视频镜头、作业三的游戏素材都适用。

Revisit: L02 · 9 (The edit contract), L02 · 10 (An invariant can be semantic or exact)

12. Choose a control by the information that is missing

Here's Lecture 2's rule as a table. Text says what should change. A mask says where change is allowed, though it won't fix seams or reflections on its own. A spatial condition, meaning edges, depth or pose through ControlNet, says what shape or structure to follow. ControlNet keeps the pretrained path frozen and adds a trainable copy joined by zero-initialized convolutions. A reference image, through IP-Adapter, carries appearance that words can't describe precisely, instantly but less faithfully. And when a concept has to recur across many shots, you adapt the weights with LoRA or DreamBooth. The mistake to avoid is adding adjectives when the missing information is spatial.

第 2 讲的规则做成表。文字说明改什么；遮罩说明哪里允许改，但它本身不解决接缝和反射；空间条件（ControlNet 的边缘、深度、姿态）说明要遵循的形状结构，ControlNet 冻结原模型，加一个由零初始化卷积连接的可训练副本；参考图（IP-Adapter）传达文字说不清的外观，即时但不够忠实；概念需要在多个镜头中反复出现时，就用 LoRA 或 DreamBooth 调整权重。要避免的错误：缺的是空间信息，却还在加形容词。

Revisit: L02 · 29 (Different controls carry different information), L02 · 30 (ControlNet: add spatial conditioning), L05 · 64 (Image Prompt Adapter (IP-Adapter))

13. Inside an image editor: the target is missing at inference

Lecture 2 opened a current image editor from the input side. The source image enters twice: semantic features through a vision-language component, which help with what the pear is and what "it" refers to, and visual latents through the VAE, which keep the particular shape and texture recognizable. The target stream starts as noise, and

a transformer processes it with those conditions. The key point: during training we have the desired edited target, so we can measure the error. At inference we have the source and the request, but the target doesn't exist yet. Showing an editor a reference conditions one generation. It doesn't teach new weights; that's adaptation, like LoRA. Where does instruction-following come from? Training triples of source, instruction and target, starting with InstructPix2Pix's synthetic data. And if a pair says "make it ceramic" but the target also moves the camera, the model learns that unwanted association. The preservation contract starts in the training examples.

第 2 讲从输入端打开一个当前的图像编辑器。源图进入两次：经视觉语言组件得到语义特征，帮助理解梨是什么、“它”指什么；经 VAE 得到视觉潜变量，保持具体形状和纹理可辨。目标流从噪声开始，Transformer 结合这些条件处理它。关键：训练时有期望的编辑结果，可以衡量误差；推理时只有源图和要求，目标尚不存在。给编辑器看参考图只是为这一次生成提供条件，不会学到新权重，那是 LoRA 这类适配做的事。遵循指令的能力从哪来？来自源图、指令、目标三元组的训练，最早是 InstructPix2Pix 的合成数据。如果一对样本说“变成陶瓷”，目标却连镜头也动了，模型就会学到这种不想要的关联。保持契约从训练样本开始。

Revisit: L02 · 17 (Inside a current image editor), L02 · 18 (The missing target at inference), L02 · 25 (Where edit supervision comes from), L02 · 26 (What a mismatched training pair teaches)

14. Count what the model actually receives

Two small calculations from Lecture 2 that catch people out. First, image latents. A 1024 by 1024 RGB image is just over three million values. With sixteen-times spatial reduction and sixty-four latent channels, it becomes 64 by 64 by 64, which is 262,144 values. Sixteen times smaller along each axis doesn't mean sixteen times fewer values, because the channel count changes too. Here it's about twelve times fewer. Second, video latents. Forty-five frames with a special first frame and four-times temporal compression: forty-five is four k plus one, so k is eleven and there are twelve latent frames, which fit a three-by-four virtual grid. Simply dividing forty-five by four misses the first-frame convention. And one more piece from Lecture 2's appendix: InstructPix2Pix has two guidance scales, one for the image and one for the text. Set both to one and the middle terms cancel, leaving the fully conditioned prediction.

第 2 讲两个容易出错的小计算。一，图像潜变量：1024×1024 的 RGB 图约三百多万个值；空间缩小 16 倍、64 个潜通道后变成 64×64×64，即 262,144 个值。每个轴缩小 16 倍不等于值的数量少 16 倍，因为通道数也变了，这里约少 12 倍。二，视频潜变量：45 帧，首帧特殊处理，时间压缩 4 倍：45=4k+1，k=11，所以有 12 个潜帧，正好排成 3×4 的虚拟网格，直接用 45 除以 4 会漏掉首帧约定。附录里还有一点：InstructPix2Pix 有图像和文本两个引导尺度，两者都设为 1 时中间项相消，剩下完全条件化的预测。

Revisit: L02 · 13 (A smaller workspace: image latents), L02 · 85 (45 frames become 12: why?), L02 · 115 (Separate image and text guidance)

15. Art direction: a reference is a relationship, not a label

The middle of Lecture 2 was pure art direction. The brief for AFTER RAIN said "fragile", and we asked how to make fragility visible. Scale, restrained light and a reflection less stable than the object are hypotheses you can test, not a formula. Then Hokusai's Great Wave. Look for Mount Fuji: it's small and distant, and the wave's rhythm creates the tension. We didn't borrow "Hokusai style". We borrowed a relationship: a small, vulnerable form facing a much larger environment. Then reference roles: one image for the subject, one for the composition, one for the material, and you check for unwanted transfers by removing one reference and seeing what disappears. And negative space has a job: it makes the object feel small and gives the title somewhere to live. If you reference an artist in your final presentation, say which relationship you borrowed.

第 2 讲中段是纯粹的艺术指导。AFTER RAIN 的简报写着“脆弱”，我们问如何让脆弱可见：尺度、克制的光、比物体更不稳定的倒影，这些是可以检验的假设，不是公式。然后是北斋的《神奈川冲浪里》：找富士山，它又小又远，浪的节奏制造张力。我们借的不是“北斋风格”，而是一种关系：小而脆弱的形体面对大得多的环境。参考的角色：一张管主体、一张管构图、一张管材质，去掉一张看看什么消失了，以此检查不想要的迁移。负空间也有作用：让物体显得小，并给标题留位置。期末展示若引用艺术家，请说明你借的是哪种关系。

Revisit: L02 · 51 (Hokusai: scale, rhythm, and tension), L02 · 52 (Reference as a relationship, not a label), L02 · 57 (Reference roles), L02 · 54 (Negative space has a job)

16. A lucky image or a reliable direction?

Two more habits from Lecture 2 that I expect to see in your final presentations. First: a lucky image or a reliable direction? You compare two prompts once and the second gives a wonderful poster. Was it the wording, or a favourable random sample? Several outputs per condition tell you. And there are two different questions: would I exhibit this image, and could I build a series in this direction? Second: a critique with observable evidence. Name a visible feature, explain its effect, and propose a revision. For example: "the crisp reflection makes the collage feel photographic, so I'd simplify its shape." Compare that with "I don't like the reflection." The first gives the artist something to do. And a process record, with the intention, the invariants and one rejected result, makes that critique precise instead of a guess.

第 2 讲还有两个习惯，我希望在期末展示中看到。第一：幸运的一张图，还是可靠的方向？两个提示词各试一次，第二个出了一张好海报，是措辞的功劳还是随机运气？每个条件多生成几次就知道了。还有两个不同的问题：我会展出这一张吗？我能沿这个方向做一个系列吗？第二：用可观察的证据点评。说出一个可见的特征、它的效果、一个修改建议，比如“清晰的倒影让拼贴显得像照片，我会简化它的形状”，而不是“我不喜欢这个倒影”。记录意图、不变量和一个被否决的结果，能让点评更精确。

Revisit: L02 · 62 (A lucky image or a reliable direction?), L02 · 67 (A critique with observable evidence), L02 · 70 (A process record makes critique more precise)

17. A video edit must survive the next frame

Then the edit has to move. In Lecture 2's third part, a video edit has to keep its promises in the next frame. Three lessons. Consistency isn't stillness: the stem may hide during a turn, but it shouldn't change colour without a reason. Four good frames can hide a bad video, so a contact sheet is publicity stills, not the performance, and you review full playback around turns and occlusions. And the frozen video wins the wrong test: a naive frame-difference score prefers a video that never moves. That was our first meeting with a metric that rewards the wrong thing. Keep it in mind, because Thread 4 is built on it.

编辑还要动起来。第 2 讲第三部分：视频编辑要在下一帧继续兑现承诺。三点：一致不等于静止，转身时梗可以被挡住，但不该无故变色；四张好帧可以掩盖一段坏视频，联系表只是剧照，不是表演，要完整播放，重点看转身和遮挡；静止的视频赢了错误的测试，简单的帧差分数会偏爱不动的视频。这是我们第一次遇到奖励错误东西的指标，记住它，主线四就建立在这上面。

Revisit: L02 · 76 (Consistency is not stillness), L02 · 80 (Four good frames can hide a bad video), L02 · 94 (The frozen video wins the wrong test)

18. Inside the generator

Lecture 3, September 15th: inside the generator. The subtitle asked "why does it produce this?", and we answered with arithmetic you could do on paper. A traveller in a yellow coat on a long platform was our running example, because the colour, the identity and the perspective were easy to check separately. We covered noise and denoising, VAEs, CLIP and attention, transformers, flow matching, seeds, guidance and steps, LoRA and ControlNet, a look at 2026 image and video models, and the first glimpse of Astra, world models and robot policies, which came back in Lecture 9. If you remember one sentence from that night, make it this one: the model predicts, the sampler takes a step.

第 3 讲，9 月 15 日：生成器内部。副标题问“它为什么产出这个？”，我们用纸上就能算的算术来回答。示例是站台上穿黄外套的旅人，因为颜色、身份、透视都容易分别检查。内容包括加噪去噪、VAE、CLIP 与注意力、Transformer、流匹配、种子、引导、步数、LoRA 和 ControlNet、2026 年的图像和视频模型，以及对 Astra、世界模型、机器人策略的第一次介绍，它们在第 9 讲又回来了。那晚只记一句话的话：模型负责预测，采样器负责走一步。

Revisit: L03 · 1 (Inside the generator), L03 · 21 (The model predicts. The sampler takes a step.), L03 · 99 (University lectures behind this lesson)

19. Close is not correct: CLIP, relations and continuity

Lecture 3 explained why a picture can match the topic and still be wrong. CLIP learns from pictures paired with captions, so it learns associations. But captions mention the person and leave out the railing, so precise relations are weakly supervised. That's why you get a station, a traveller and a suitcase, but the suitcase is on the wrong

side, or you ask for two objects and get three. So your checks should include the relations that carry the meaning, not just a list of nouns. Then continuity: a seed is not a character description. The character sheet defines what must stay: the silhouette, the surface details like buttons and pockets, and ownership, meaning which hand holds the prop. A visible side of a suitcase should change when the camera moves. Its handle shouldn't change construction for no reason.

第 3 讲解释了为什么一张图主题对了却仍然错。CLIP 从图片和配对的说明文字中学习联想，但说明文字会提到人而漏掉栏杆，所以精确的关系监督很弱。于是你得到车站、旅人和行李箱，但箱子在错误的一侧；要两个物体却得到三个。所以检查清单应包括承载意义的关系，而不只是名词。连贯性：种子不是角色描述。角色设定图定义必须保持的东西：轮廓、纽扣口袋等表面细节、归属（哪只手拿道具）。镜头移动时箱子可见的一面应该变，但把手的结构不该无故改变。

Revisit: L03 · 33 (A similar-looking image can still be wrong), L03 · 32 (How CLIP links pictures and words), L03 · 51 (A character sheet defines what must stay)

20. Attention: every token chooses what matters

The second engine part is attention. Each token makes a query, a key and a value. Queries ask, keys match, values contribute, and the softmax weights sum to one. In Lecture 1 the example was a red scarf: a token in frame twenty can attend to the tokens that carried the scarf earlier. Cross-attention is how words reach image regions, which is why Lecture 5's Custom Diffusion only fine-tuned the key and value projections of cross-attention and still learned a new concept. In Lecture 4 we split it into spatial attention within a frame and temporal attention across frames. One warning from that lecture still matters: attention connects features, but it doesn't guarantee understanding.

第二个部件是注意力。每个 token 生成 query、key、value：query 提问，key 匹配，value 贡献，softmax 权重之和为一。第 1 讲的例子是红围巾：第 20 帧的 token 可以关注之前带着围巾的 token。交叉注意力让文字作用到图像区域，所以第 5 讲的 Custom Diffusion 只微调交叉注意力的 key 和 value，就能学会新概念。第 4 讲把它分成帧内的空间注意力和跨帧的时间注意力。那一讲的提醒仍然重要：注意力连接特征，但不保证理解。

Revisit: L01 · 26 (Queries ask; keys match; values contribute), L04 · 12 (Spatial attention and temporal attention), L05 · 34 (Only fine-tune cross-attention layers)

21. Learn to remove noise, then learn how to move

Now the core idea from Lecture 3, with the numbers we did by hand. Training adds noise to a clean image: a weighted clean value plus weighted noise. With a clean value of point eight, noise of minus point two, and weights point six and point eight, we got point three two. The network learns to predict the noise, or, in flow matching, the direction of movement from noise to data: the target minus the starting noise. At generation time there is no target. We follow the learned directions. And keep the division of labour from Lecture 3: the model

predicts, the sampler takes the step. Change the sampler and the same network can give a different picture.

第 3 讲的核心，用我们手算过的数字。训练时给干净图像加噪：干净值加权加噪声加权。干净值 0.8、噪声-0.2、权重 0.6 和 0.8，结果是 0.32。网络学会预测噪声，或者在流匹配里预测从噪声到数据的方向：目标减起点噪声。生成时没有目标，只是沿着学到的方向走。记住分工：模型负责预测，采样器负责走一步。换采样器，同一个网络也可能给出不同的图。

Revisit: L03 · 13 (Clean image plus noise), L03 · 44 (Flow matching: learning how to move), L03 · 21 (The model predicts. The sampler takes a step.)

22. Three knobs, three different jobs

These are the three knobs you touched in every assignment. The seed is an address into a repeatable random process, not an instruction. Seed forty-two doesn't mean wide shot. To repeat a result, save the whole workflow, not just the seed. Guidance pushes past the prompted prediction: unconditioned plus scale times the difference. With predictions of two and three, a scale of four gives six, outside the range. That's why strong guidance burns colours and makes samples look alike. Steps set how accurately you follow the learned field, not whether the field is right. Fifty steps won't fix a wrong hand. And a distilled model trained for one to four steps is a different system, not a slow model stopped early.

这是每次作业都碰到的三个旋钮。种子是可重复随机过程的地址，不是指令，42 号种子不代表远景；要复现结果，保存整个工作流，而不只是种子。引导会越过有条件预测：无条件加上尺度乘以差值。预测值 2 和 3，尺度 4 得到 6，超出了范围，所以强引导会让颜色过饱和、样本趋同。步数决定你沿学到的场走得多准，不决定场本身对不对，50 步修不好错的手。为一到四步蒸馏训练的模型是另一个系统，不是提前停下的慢模型。

Revisit: L03 · 48 (Different seeds can give different pictures), L03 · 54 (A small guidance calculation), L03 · 64 (Some models need only a few steps)

23. How to make a fair comparison

The end of Lecture 3 turned the knobs into an experiment. To ask whether guidance changes subject clarity and variety, keep the model, prompt, size, sampler and steps fixed, and try different guidance values on the same set of seeds. Write the comparison as a table before you run it: rows are seeds, columns are guidance values. Decide your judging questions beforehand, because if you decide what counts as success after seeing the images, you'll favour the one you happened to like. Record failures too, because leaving them out changes what the grid represents. And label your figures: a progress picture from partway through one run answers a different question from finished runs at fifteen and fifty steps. This is exactly what the reconstruction log in Assignment 2 asks for: at least three attempts, one change each.

第 3 讲结尾把旋钮变成实验。要问引导是否改变主体清晰度和多样性：固定模型、提示词、尺寸、采样器

和步数，在同一组种子上试不同的引导值。运行前先把对比写成表：行是种子，列是引导值。事先决定评判问题，否则看到图后再定义成功，就会偏爱碰巧喜欢的那张。也记录失败，因为省略失败会改变网格的含义。给图加标签：一次运行中途的进度图和 15 步、50 步的完成图回答的是不同问题。这正是作业二重建日志的要求：至少三次尝试，每次只改一处。

Revisit: L03 · 91 (How to make a fair comparison), L03 · 94 (Reading a guidance grid), L03 · 62 (A progress picture or a finished result?)

24. Diagnose it: which knob, or no knob at all?

Pen out. Four symptoms from real student logs this term. One: same prompt, but every run gives a completely different composition. Two: the subject is unmistakable, but colours are burnt and every sample looks the same. Three: the result is fine, but it takes forty seconds and you need it live in a gallery. Four: a character's face changes after she turns away and turns back. For each one, write down which knob you'd touch, or whether it isn't a knob at all. Thirty seconds. Then compare with your neighbour.

拿出笔。四个症状，来自本学期同学的记录。一：同一个提示词，每次构图完全不同。二：主体非常明显，但颜色过饱和，每个样本都一样。三：效果可以，但要 40 秒，而你需要在展厅里实时运行。四：角色转身再转回来，脸变了。每个写下你会调哪个旋钮，或者它根本不是旋钮问题。三十秒，然后和旁边同学对一下。

Revisit: L03 · 58 (Discussion 3: Clear subject, wrong mood), L07 · 22 (Test it yourself: does she come back the same?)

25. Diagnosis: three knobs and one memory problem

Here are the answers. A is the seed: fix it for comparisons and record the whole workflow, and use references when you need a particular layout. B is guidance set too high: lower the scale and compare across the same list of seeds, changing one variable at a time, as Lecture 1 asked. C is cost: a distilled model, fewer steps, quantization from Yaofu's talk, or a smaller resolution, and you measure seconds, not step counts. D is the trick question. It isn't a knob. It's memory over time: the representation and the context window. We'll meet it again in Lectures 7 and 9, and after Lecture 9 you'll see it's one problem wearing three costumes.

答案。A 是种子：比较时固定，记录完整工作流，需要特定构图就用参考图。B 是引导太高：降低尺度，在同一组种子上比较，一次只改一个变量。C 是成本：蒸馏模型、更少步数、Yaofu 讲的量化或降低分辨率，要测秒数而不是步数。D 是陷阱题：不是旋钮，是时间上的记忆，即表示和上下文窗口。第 7、9 讲还会遇到它，第 9 讲之后你会看到它是同一个问题的三种装扮。

Revisit: L03 · 49 (How to repeat a result), L03 · 63 (Generation cost: count model evaluations), L09 · 19 (Under the hood: why the world must come back the same)

26. Video generation and art filmmaking

Lecture 4, September 22nd: generative AI for art, images and video. It was built around five readings and a running example, an empty studio with a painting on the wall, and it moved through five areas: video generation models, visual art, filmmaking, visual computing, and the creative industries. The question was: what changes when an image moves? A still suggests possibilities, but a video commits to a sequence of events. If a hand rises, it has to belong to the same person as it moves. This was also the lecture where the art students' questions and the technical students' questions started to be the same questions.

第 4 讲，9 月 22 日：用于艺术、图像和视频的生成式 AI。围绕五篇读物和一个示例：墙上挂着一幅画的空工作室。内容覆盖五个领域：视频生成模型、视觉艺术、电影制作、视觉计算、创意产业。问题是：图像动起来时什么改变了？静态图暗示可能性，视频则承诺一连串事件，一只手抬起，它在运动中必须属于同一个人。也是从这一讲起，艺术学生和技术学生的问题开始变成同一个问题。

Revisit: L04 · 1 (Generative AI for art, images, and video), L04 · 9 (A video has two kinds of time), L04 · 70 (A short film still needs decisions)

27. Two kinds of time, two kinds of motion

Three ideas from Lecture 4's first part that you need to read any video paper. First, two kinds of time. Frame time is the time inside the scene: a person enters, crosses the studio and looks at the painting. Denoising time is the sequence of computational steps that produces the sample. A denoising step isn't an extra frame. Whenever you see a time index, ask which one it is. Second, spatial attention relates things within a frame, such as the hand and the object it holds, and temporal attention relates the hand now to the hand a moment later. Third, object motion and camera motion are different. A person can cross a still frame, or the camera can circle a still person, and both change pixel positions. If you want to control them separately, you have to specify them separately.

第 4 讲第一部分的三个想法，读任何视频论文都需要。一，两种时间：画面时间是场景内的时间，有人进来、穿过工作室、看那幅画；去噪时间是生成样本的计算步骤，一个去噪步不是多出来的一帧。看到时间下标就问是哪一种。二，空间注意力关联同一帧里的事物，比如手和它拿的东西；时间注意力关联此刻的手和下一刻的手。三，物体运动和镜头运动不同：人穿过静止画面，或镜头绕着静止的人转，都会改变像素位置。想分别控制，就要分别说明。

Revisit: L04 · 9 (A video has two kinds of time), L04 · 12 (Spatial attention and temporal attention), L04 · 20 (Motion and camera control)

28. Data shapes what a model can make

Then data. Lecture 4 showed a pipeline that selects clips, filters them and attaches captions, and every stage changes what reaches training. Filtering trades coverage for usable examples: removing difficult material may improve average quality while leaving gaps in certain actions or styles. If your project depends on an unusual

visual language and the model keeps resisting it, rewriting the prompt may not help, because the gap may be in the data. And here's what filmmakers said they wanted in the survey from that night: character identity, body movement, camera control and local editing. All four showed a gap between how important artists rated them and how well tools performed. That chart predicted much of the semester: Lecture 5 on identity, Lecture 7 on movement and dance, Lecture 8 on performance.

然后是数据。第 4 讲展示了一个挑选片段、筛选、加说明文字的流程，每一步都改变进入训练的东西。筛选用覆盖面换取可用样本：去掉困难素材可能提升平均质量，却在某些动作或风格上留下空白。如果你的项目依赖不寻常的视觉语言而模型一直抗拒，改写提示词可能无济于事，因为空白可能在数据里。那晚调查中电影人最想要的：角色身份、身体运动、镜头控制、局部编辑，四项都显示重要性和工具表现之间有差距。这张图预告了本学期的大部分内容：第 5 讲讲身份，第 7 讲讲运动和舞蹈，第 8 讲讲表演。

Revisit: L04 · 15 (Training data shapes the result), L04 · 16 (A closer look at data filtering), L04 · 56 (What filmmakers wanted)

29. Intention first: one intention, one intervention

The visual-art part of Lecture 4 started from the artist, not the tool. An intention becomes useful when you connect it to visible decisions. Suppose the intention is an uneasy memory: delayed movement, missing details, or a familiar room with one altered object. Each suggests different controls, while "make it more artistic" suggests nothing. A scenario gives the work a purpose: the same image could be a print, a storyboard frame, or one frame of an installation. Content and style need separate decisions: if a character must stay recognizable while the image becomes painted, decide which features stay, such as silhouette, proportions and pose, and which may vary, such as brushwork and palette. The exercise was one intention and one visible intervention, plus a way to check whether the audience reads it.

第 4 讲视觉艺术部分从艺术家出发，而不是从工具出发。意图只有和可见的决定联系起来才有用。比如意图是不安的记忆：延迟的动作、缺失的细节、熟悉的房间里一件变了的物体。每个选择对应不同的控制，而“更艺术一点”什么也没说。情境给作品目的：同一张图可能是版画、分镜或装置中的一帧。内容和风格要分开决定：角色要保持可辨认、画面变成绘画风格时，决定哪些保持（轮廓、比例、姿态）、哪些可以变（笔触、色调）。练习是：一个意图、一个可见的干预，再加一个检验观众是否读懂的方法。

Revisit: L04 · 37 (The artist’s intention comes first), L04 · 43 (Content and style need separate decisions), L04 · 49 (One intention, one intervention)

30. Five art films, five different workflows

The filmmaking part of Lecture 4 looked at real art films from a 2025 survey. The headline: the films used on average 3.14 video-generation tools each, because different stages have different needs. Overthinking built a nostalgic visual language through image adaptation, asset creation, animation and editing. Clown turned inconsistency into expression: frame-by-frame style change supporting a fragmented identity. That's a deliberate

choice, and an accidental change doesn't automatically mean anything. For Pixi kept a recurring character through iterative generation and region editing. Fish Tank went through 3D, with manual UV mapping and topology, ordinary production work that doesn't disappear. And volumography captured space, so the camera became a decision made after capture, which led straight to splats in Lecture 9. The common lesson: a short film still needs decisions about what changes, when, and what the audience should feel.

第 4 讲电影部分看了 2025 年调查中的真实艺术短片。要点：每部片平均使用 3.14 个视频生成工具，因为不同阶段需求不同。《Overthinking》通过图像改编、素材制作、动画和剪辑建立怀旧的视觉语言。《Clown》把不一致变成表达：逐帧风格变化支持碎片化的身份，这是有意的选择，偶然的变化不会自动有意义。《For Pixi》通过迭代生成和区域编辑保持一个反复出现的角色。《Fish Tank》走 3D 路线，手工做 UV 和拓扑，常规的制作工作并不会消失。体积摄影捕捉空间，让镜头成为拍摄后的决定，直接通向第 9 讲的高斯泼溅。共同的一课：短片仍需决定什么改变、何时改变、观众应该感受什么。

Revisit: L04 · 58 (Overthinking), L04 · 61 (Clown), L04 · 64 (For Pixi), L04 · 66 (Fish Tank), L04 · 68 (Volumography)

31. Quality has several meanings

Lecture 4 ended with a warning that became Lecture 10. A technical score answers a particular question: prompt alignment, visual similarity or some aspect of perceptual quality. It can't establish whether an artwork succeeds. A deliberately ambiguous image may score badly on literal prompt matching and still produce exactly the intended experience. Quality has several meanings: an image can be sharp and fail to convey the action, and a sequence can match the prompt and feel emotionally wrong. So make separate judgments: technical continuity, clarity of the reveal, and the audience's interpretation. And the last slide: a process others can understand. Save references and meaningful before-and-after versions, record models and settings, and explain the decisions that changed the work, including editing and sound. That's the same advice as Lecture 7's project folder and Lecture 10's credit block, a month earlier.

第 4 讲结尾的提醒后来变成了第 10 讲。技术分数只回答一个特定问题：提示词对齐、视觉相似度或某方面的感知质量，它无法确定一件艺术作品是否成功。刻意模糊的图在字面匹配上得分低，却可能正好产生想要的体验。质量有多种含义：图像清晰却没传达动作，序列符合提示词却情感不对。所以分开判断：技术连贯、揭示是否清楚、观众如何解读。最后一页：让别人能理解的过程。保存参考和有意义的前后版本，记录模型和设置，解释改变作品的决定，包括剪辑和声音。这和第 7 讲的项目文件夹、第 10 讲的署名块是同一条建议，只是早了一个月。

Revisit: L04 · 44 (A technical score answers a narrow question), L04 · 98 (Quality has several meanings), L04 · 99 (A process others can understand)

32. Making the model yours

Lecture 5, September 29th: making the model yours, with two tracks running together. The technical track covered Textual Inversion, DreamBooth, Custom Diffusion, concept ablation, LoRA, adapters and speed. The art

track covered copyright, memorization, living artists, provenance and bias. Many of the slides were adapted from Jun-Yan Zhu's course at CMU, including the moongate and his dog Stark. The question opened and closed the evening: whose model is it? Keep your answer from that night in mind, because I'll ask you about it at the end of this section.

第 5 讲，9 月 29 日：把模型变成你的，两条线并行。技术线：Textual Inversion、DreamBooth、Custom Diffusion、概念消除、LoRA、适配器和速度。艺术线：版权、记忆化、在世艺术家、溯源和偏见。很多页改编自 CMU 朱俊彦的课程，包括月洞门和他的狗 Stark。问题贯穿那晚的开头和结尾：模型是谁的？记住你那晚的答案，这一部分结束时我会再问。

Revisit: L05 · 1 (Making the Model Yours), L05 · 12 (Customization: V* dog), L05 · 78 (Whose model is it — now?)

33. What prompting alone cannot fix

Why customize at all? Lecture 5 named two bottlenecks. The first is linguistic: you can't describe the exact line quality of your own drawing, or a friend's face, in words. The second is data: personal things, like your dog or your family's shop, aren't in the training set. The moongate was the example. Ask Stable Diffusion for a moongate, the circular gateway you can find in Hong Kong gardens, and you get something plausible but wrong. The model doesn't know the concept. Then Stark, Jun-Yan's dog. Ask for a dark grey weimaraner and you get a perfectly good weimaraner, but never Stark. A description isn't an identity, and that's exactly the character drift many of you fought in Assignment 1. After customization, "V-star dog wearing sunglasses" gives you Stark, in sunglasses, and a moongate in snow is something no training photo showed.

为什么要定制？第 5 讲指出两个瓶颈。一是语言：你无法用文字描述自己画的线条质感或朋友的脸。二是数据：私人的东西，比如你的狗、家里的店，不在训练集里。例子是月洞门：让 Stable Diffusion 画月洞门（香港园林里就有的圆形门洞），结果看似合理却是错的，模型不懂这个概念。然后是朱俊彦的狗 Stark：要深灰色魏玛犬，你会得到一只很好的魏玛犬，但永远不是 Stark。描述不等于身份，这正是很多人在作业一中对抗的角色漂移。定制后，“V*狗戴墨镜”就是戴墨镜的 Stark，雪中的月洞门也是任何训练照片里都没有的。

Revisit: L05 · 7 (Text-to-image isn't perfect: the moongate), L05 · 11 (No knowledge of personal concepts), L05 · 12 (Customization: V* dog), L05 · 10 (Unseen contexts)

34. Making the model yours: what actually gets trained?

Lecture 5 in one table: the key question is what gets trained. Textual Inversion keeps the whole model frozen and learns one new word, an embedding. It's good for styles and weak on identity. DreamBooth fine-tunes all the weights and preserves identity best, but it overfits, so you regularize with images of the general class. Custom Diffusion trains only the key and value projections in cross-attention, so the update is small and concepts combine. LoRA adds a thin overlay of two skinny matrices: for a 4096 by 4096 layer at rank sixteen, that's about 131 thousand numbers instead of 16.8 million. IP-Adapter trains nothing per subject; an encoder turns your

reference into a visual prompt. The rule of thumb from that night: a style can live in a word, but a character who must stay the same across shots needs a change to the weights.

第 5 讲一张表：关键是训练了什么。Textual Inversion 冻结整个模型，只学一个新词（嵌入），适合风格，身份保持弱。DreamBooth 微调全部权重，身份保持最好，但会过拟合，需要用同类别图像正则化。Custom Diffusion 只训练交叉注意力的 key 和 value，更新小，概念能组合。LoRA 加一层由两个窄矩阵组成的薄层：4096×4096 的层、秩 16，约 13.1 万个数，而不是 1678 万。IP-Adapter 不针对主体训练，编码器把参考图变成视觉提示。那晚的经验：风格可以放在一个词里，但要跨镜头保持同一个角色，就需要改权重。

Revisit: L05 · 27 (DreamBooth vs. Textual Inversion), L05 · 61 (Low-rank adaptation (LoRA)), L02 · 35 (LoRA: low-rank adaptation)

35. Overfit, memorize, forget

The same mechanism has three faces. Overfit: after learning "moongate", the fine-tuned model puts moongates into every image of the moon. The fix is regularization images, real photos of related concepts, because what you leave out of a dataset teaches as much as what you put in. Memorize: at internet scale, Carlini and colleagues generated from the captions of heavily duplicated training images, and some outputs matched the originals almost exactly. A living artist's style can be memorized too, which is why "in the style of Greg Rutkowski" became a famous prompt. Forget: concept ablation, also from Kumari's group, makes the model map "Grumpy Cat" to an ordinary cat and remove the Rutkowski look while keeping the boat and the water. That's what an opt-out inside the model could look like. For your own small datasets: remove duplicates and include what the model should keep.

同一机制有三张面孔。过拟合：学会“月洞门”后，微调的模型在每张月亮图里都放上月洞门，解决办法是正则化图像，即相关概念的真实照片，因为你从数据集中排除的东西和放进去的一样有教育作用。记忆化：在互联网规模上，Carlini 等人用重复多次的训练图像的说明文字生成，有些结果几乎和原图一样；在世艺术家的风格也会被记住，所以“Greg Rutkowski 风格”成了著名提示词。遗忘：同样来自 Kumari 团队的概念消除，让模型把“Grumpy Cat”变成普通猫，去掉 Rutkowski 的风格但保留船和水，这就是模型内部的退出机制的样子。对你自己的小数据集：去重，并包含模型应保留的东西。

Revisit: L05 · 35 (How to prevent overfitting?), L05 · 28 (Memorized training images), L05 · 54 (Ablating Greg Rutkowski (1)), L05 · 45 (Concept ablation: the goal)

36. LoRA in one line: a thin overlay on a frozen layer

The arithmetic behind the most common customization method, which came up in Lecture 2 and was the centre of Lecture 5. A layer's weight matrix W stays frozen. LoRA learns an update that's the product of two skinny matrices, B times A, with rank r. For a 4096 by 4096 layer, a full update has 16,777,216 entries. At rank sixteen, the two factors together have two times 4096 times 16, which is 131,072, a factor of 128 fewer for that update. That's why a LoRA file is small and you can swap overlays on the same base model. Two cautions from those

nights. The base weights still exist, so the whole model didn't shrink by 128. And a small update doesn't prove the concept generalizes: test it on poses and settings the training images didn't show. Custom Diffusion makes a similar bet differently, training only the key and value projections of cross-attention.

最常用定制方法背后的算术，第 2 讲提到、第 5 讲重点讲过。层的权重矩阵 W 冻结，LoRA 学一个由两个窄矩阵 B×A 构成的更新，秩为 r。4096×4096 的层完整更新有 16,777,216 个数；秩 16 时两个因子共 2×4096×16=131,072 个，这一更新少了 128 倍。所以 LoRA 文件很小，可以在同一个基础模型上换不同的叠加层。两个提醒：基础权重仍在，整个模型并没有缩小 128 倍；更新小不代表概念能泛化，要在训练图里没有的姿态和场景中测试。Custom Diffusion 用另一种方式下同样的注：只训练交叉注意力的 key 和 value 投影。

Revisit: L05 · 61 (Low-rank adaptation (LoRA)), L02 · 35 (LoRA: low-rank adaptation), L05 · 34 (Only fine-tune cross-attention layers)

37. Defaults have politics

Lecture 5 ended with bias, and it belongs in a lecture about customization, because every base model's defaults come from someone else's dataset. PULSE, in 2020, took a pixelated face and searched a generative model for a sharp face that matched. Where do the details come from? Not from the input: they come from the model's prior, and for a well-known pixelated photo of Barack Obama, PULSE produced a white face. Then the truncation trick: a setting that improves quality by pulling samples toward the average also reduces diversity, and the majority group grows as truncation increases. A technical setting turns out to have a social effect. In text-to-image models, "manager" gave mostly white men in suits. Quick fixes, like rewriting prompts behind the scenes, help. For artists, the longer-term answer may partly lie in building your own data and shaping your own models.

第 5 讲以偏见结尾，它属于定制这一讲，因为每个基础模型的默认都来自别人的数据集。2020 年的 PULSE 把像素化的人脸放进生成模型里搜索匹配的清晰人脸。细节从哪来？不是输入，而是模型的先验；对一张著名的奥巴马像素照，PULSE 生成了一张白人的脸。截断技巧：把样本拉向平均来提升质量，同时也降低多样性，截断越强，多数群体占比越大。一个技术设置原来有社会效果。文生图模型里，“经理”大多是穿西装的白人男性。幕后改写提示词之类的快速修补有帮助；对艺术家来说，更长远的答案部分在于建立自己的数据、塑造自己的模型。

Revisit: L05 · 71 (Danger and ethical concerns: PULSE (1)), L05 · 74 (The truncation trick reduces diversity), L05 · 75 (Bias in text-to-image models), L05 · 76 (Quick fixes or long-term solutions?)

38. Whose model is it?

Lecture 5 began with a show of hands, and I asked you to remember where you put your hand. A foundation model has seen billions of images by millions of people. You add twenty of yours and fine-tune it. Whose work is the output: the company's, yours, or the people in the training data? That night we saw how memorized images are found, how a living artist's style becomes a prompt word, and how concept ablation can make a model forget

Greg Rutkowski while keeping the boat and the water. Hold your answer. We'll come back to it in Thread 4 with consent, and you'll need it for your credit block.

第 5 讲开头举过手，我请你们记住手举在哪里。基础模型看过数百万人的数十亿张图，你加二十张自己的图微调，输出是谁的作品：公司的、你的，还是训练数据里那些人的？那晚我们看到如何发现被记忆的图像、在世艺术家的风格如何变成提示词、概念消除如何让模型忘掉 Greg Rutkowski 却保留船和水。先留着你的答案，主线四讲同意时再回来，你的署名块也需要它。

Revisit: L05 · 3 (Whose model is it?), L05 · 30 (Memorized style), L05 · 54 (Ablating Greg Rutkowski (1))

39. Week 6: store less, so you can run it

Week 6 was Yaofu's talk on quantization for faster model inference. There were no slides on our site, so here's how it kept coming back in later lectures, which is the best measure of a good lecture. The idea: keep each number in the model with fewer bits, like saving a photo as a smaller JPEG. Lecture 7 did the arithmetic for a fourteen-billion-parameter model like Helios or Wan-Dancer: at sixteen bits, two bytes per parameter, that's twenty-eight gigabytes for the weights alone, fourteen at eight bits, seven at four bits, before activations, the VAE and the text encoder. That's the difference between a datacentre GPU and one you might own. And Lecture 5 put it in a family of three: adapters change less, distillation steps less, quantization stores less. Lecture 8 added that real-time avatars use all three at once.

第 6 周是 Yaofu 关于量化加速推理的讲座，网站上没有幻灯片，所以这里看它在之后各讲中如何反复出现，这是衡量一讲好坏的最好方式。思路：模型里每个数用更少的比特保存，就像把照片存成更小的 JPEG。第 7 讲为 Helios、Wan-Dancer 这样 140 亿参数的模型算过：16 位时每参数 2 字节，仅权重就是 28GB；8 位 14GB；4 位 7GB，还不算激活、VAE 和文本编码器。这就是数据中心 GPU 和你可能拥有的 GPU 之间的差别。第 5 讲把它放进三兄弟：适配器改得少，蒸馏走得少，量化存得少。第 8 讲补充：实时数字人三样同时用。

Revisit: L07 · 61 (Under the hood: can you run an open model?), L05 · 67 (Smaller changes, fewer steps, smaller numbers), L08 · 65 (Three tricks you already know)

40. Music, minutes, and machines

Lecture 7, October 13th: what's new in AI motion. From this lecture on, the format changed to four stories from this year, each with an "under the hood" section adapted from courses at Stanford and CMU, a question for the room, and a debate at the end. The four stories: the five-second clip is over; music becomes the choreographer; motion leaves the screen; and Sora is gone. The question underneath was what happens when generation gets long, fast and physical. It was also the night Assignment 2's project-folder advice was born.

第 7 讲，10 月 13 日：AI 运动有什么新进展。从这一讲起格式变成今年的四个故事，每个都有改编自斯坦福和 CMU 课程的“底层原理”、一个给全班的问题，最后是辩论。四个故事：五秒片段的时代结束了；音乐成为编舞；运动离开屏幕；Sora 消失了。底下的问题是：当生成变得更长、更快、更有物理性时会发生什

么。作业二的项目文件夹建议也诞生于这一晚。

Revisit: L07 · 1 (What's New in AI Motion), L07 · 16 (Helios: a minute of video, in real time), L07 · 59 (Eighteen months as a product)

41. From the shot to the scene

Lecture 7 in four numbers. In 2024 a typical clip was about five seconds. Early this year Seedance 2.0 reached about fifteen. By summer, Seedance 2.5, Wan 3.0 and Kling 4.0 were at thirty seconds in one pass, long enough for an entrance, an exchange and an exit. And Helios, an open research model, made about a minute at about nineteen and a half frames per second on one H100, close to real time. There were two routes: denoise all frames at once, which is consistent but expensive, or write forward in time, autoregressively, like a language model writes words, which can stream but has to fight drift. The question we asked was what the director's job becomes when one generation is a whole scene.

第 7 讲的四个数字：2024 年典型片段约 5 秒；今年初 Seedance 2.0 约 15 秒；到夏天，Seedance 2.5、Wan 3.0、Kling 4.0 一次生成 30 秒，够一个进场、一段交流、一个退场；开源研究模型 Helios 在一块 H100 上以约每秒 19.5 帧生成约一分钟，接近实时。两条路线：所有帧一起去噪，一致但贵；或者像语言模型写字一样按时间向前自回归，可以流式但要对抗漂移。我们当时问：一次生成就是一整场戏时，导演做什么？

Revisit: L07 · 7 (The unit of AI video is now the scene), L07 · 10 (Two ways to make a long video), L07 · 16 (Helios: a minute of video, in real time)

42. How long video got cheap: write forward, cache, distil

Here's the under-the-hood part of Story 1. Why was five seconds the wall? At twenty-four frames per second, thirty seconds is six times the frames, and attention compares everything with everything, so roughly thirty-six times the comparisons. Three ideas broke the wall. Autoregression: write the video forward in chunks, the way a language model writes words, so frames can stream. The KV cache: store each layer's keys and values from earlier chunks so the past isn't recomputed. If the cache stays bounded or compressed, as in Helios, each new chunk costs about the same no matter how long the video gets. And distillation: train a student to jump to where the teacher would end up, through consistency models or DMD. Helios's distilled version uses three steps instead of about fifty. Each of these came back in Lecture 8's avatars and Lecture 9's world models.

这是故事一的底层原理。为什么五秒是一道墙？每秒 24 帧，30 秒就是六倍的帧，注意力两两比较，比较次数大约三十六倍。三个想法打破了这道墙。自回归：像语言模型写字一样按块向前写视频，帧可以流式输出。KV 缓存：保存之前各块每一层的 key 和 value，不重算过去；缓存有界或压缩（如 Helios），每个新块的成本就和视频长度无关。蒸馏：训练学生直接跳到老师会到的地方，用一致性模型或 DMD；Helios 蒸馏版只用 3 步，而不是约 50 步。这几个想法在第 8 讲的数字人和第 9 讲的世界模型中都回来了。

Revisit: L07 · 15 (Under the hood: the KV cache), L07 · 9 (Why five seconds was the limit), L07 · 10 (Two ways to

make a long video), L07 · 18 (Under the hood: distillation, from 50 steps to 3)

43. Music becomes the choreographer

Story 2. Wan-Dancer, released in July by Alibaba's Tongyi Lab with open weights. Give it a music track, one photo of the dancer and a style word, and it makes more than a minute of dance at 720p that hits the beat, in five styles: Chinese classical, K-pop, street, tap and Latin. The one idea worth remembering is that it works at two levels. The global level plans the whole song's shape: intro, verse, build, chorus and outro. The local level renders the steps beat by beat. That's how a human choreographer works, and it's why the dance stays coherent for a minute. To dance, a model has to hear beat, tempo, phrases and energy. The limits: one photo means the back of the costume is invented, and we can't check whose dances were in the training videos. Who's the choreographer here?

故事二。Wan-Dancer，7 月由阿里通义实验室以开源权重发布。给它一段音乐、一张舞者照片、一个风格词，它生成一分钟以上、720p、踩准节拍的舞蹈，五种风格：中国古典舞、K-pop、街舞、踢踏舞、拉丁舞。值得记住的一个想法是两个层级：全局层规划整首歌的结构，前奏、主歌、铺垫、副歌、尾声；局部层逐拍渲染舞步。这就是人类编舞的方式，所以舞蹈能连贯一分钟。要跳舞，模型必须听懂节拍、速度、乐句和能量。局限：只有一张照片，服装背面是编出来的；我们也无法核实训练视频里是谁的舞蹈。这里的编舞是谁？

Revisit: L07 · 34 (The trick: plan the song, then the steps), L07 · 31 (Wan-Dancer: song in, dance out), L07 · 36 (What the model has to hear), L07 · 43 (Who is the choreographer here?)

44. Motion leaves the screen

Story 3: a humanoid learned a K-pop dance challenge from phone video. There were four steps. Film a dancer. Lift the 3D joint motion from flat video. Practise in simulation with reinforcement learning, because a robot that copies a human pose exactly will fall over: it's 1.3 metres tall with 23 joints, different proportions and slower motors, so it needs retargeting and practice. Then transfer to the real robot across the sim-to-real gap, where friction, motor delays and flex never match the simulator, and domain randomization helps. Why dance challenges? They're short, repeated by thousands of people, on the beat, and filmed full body, which makes them almost perfect training data. The big idea: video is becoming training data for bodies. A dance you post now has two audiences, people and machines. We asked whether a generated video that teaches a real body to move is still "fake".

故事三：人形机器人从手机视频学会了 K-pop 舞蹈挑战。四步：拍舞者；从平面视频中提取 3D 关节运动；用强化学习在仿真中练习，因为精确照搬人类姿势的机器人会摔倒，它 1.3 米高、23 个关节、比例不同、电机更慢，需要重定向和练习；再跨越仿真到现实的差距迁移到真机，摩擦、电机延迟、形变从不和仿真一致，域随机化有帮助。为什么是舞蹈挑战？短、被成千上万人重复、踩拍、全身入镜，几乎是完美的训练数据。大想法：视频正在成为身体的训练数据，你发的舞蹈现在有两种观众：人和机器。我们问过：能教会真实身体动作的生成视频，还是“假的”吗？

Revisit: L07 · 47 (Four steps from phone to robot), L07 · 49 (Why not just copy the pose?), L07 · 52 (The sim-to-real gap), L07 · 54 (Why dance challenges are perfect training data)

45. Tools vanish. Keep your sources.

The last stop in this thread is about time in a different sense: how long your work survives. Sora became a product in December 2024. OpenAI announced its shutdown in March, the app closed in April, and the API was switched off on September 24th. It lasted about eighteen months as a product. So, from Lecture 7: keep your sources, not just your outputs. One folder per project, with a README that gives the model, version, date and settings, plus folders for prompts, references with credits, raw generations including the interesting failures, and the edit. Pin versions: the exact model ID and, for open weights, the commit hash. This isn't busywork. It's half your A2 report, and it's the evidence for your final presentation.

这条主线的最后一站，讲另一种意义上的时间：你的作品能活多久。Sora 于 2024 年 12 月成为产品，OpenAI 今年 3 月宣布关闭，4 月关闭应用，9 月 24 日关闭 API，作为产品存在约十八个月。所以第 7 讲说：保存来源，而不只是输出。每个项目一个文件夹：README 写模型、版本、日期、设置；prompts、带署名的 refs、包括有趣失败的 raw、edit 文件夹。锁定版本：准确的模型 ID，开源权重记下 commit 哈希。这不是杂务，它是作业二报告的一半，也是期末展示的证据。

Revisit: L07 · 59 (Eighteen months as a product), L07 · 65 (A project folder that survives), L07 · 66 (Useful: pin your versions)

46. When AI video learned to talk

Lecture 8, October 20th: when AI video learned to talk. It opened with a clip played with the sound off, then on. One model had made the picture, the voices, the effects and the music. The four stories: sound is no longer post-production; directing a performance, not a prompt; characters that listen; and who is the performer? Underneath was the idea from film theory that sound is half the picture. It was also the most legal and ethical lecture until Lecture 10: likeness, voice, the new SAG-AFTRA contract, and synthetic performers.

第 8 讲，10 月 20 日：AI 视频学会了说话。开场先静音播放一段片子，再开声音：画面、人声、音效和音乐都由一个模型生成。四个故事：声音不再是后期；导演表演而不是提示词；会倾听的角色；表演者是谁？底层是电影理论中“声音是画面的一半”的想法。这也是第 10 讲之前最涉及法律和伦理的一讲：肖像、声音、新的 SAG-AFTRA 合同、合成表演者。

Revisit: L08 · 1 (When AI Video Learned to Talk), L08 · 19 (Inside LTX-2: two streams, one clock), L08 · 85 (Tilly Norwood: an “actress” who never existed)

47. Sound shares the clock

Lecture 8 started with film theory. Michel Chion's synchresis: when a sound and an image happen at the same

instant, your brain welds them into one event. That's why Foley works and why bad lip-sync hurts. Then the architecture. There are four ways to wire sound and picture: picture first, sound first, two towers, or one sequence. LTX-2 uses two towers: a fourteen-billion-parameter video stream and a five-billion audio stream that look at each other in every block using only time positions, so a token at two seconds attends to audio at two seconds. That shared clock is the whole trick. And the fastest human quality check: watch p, b and m, where the lips must close.

第 8 讲从电影理论开始：Michel Chion 的“同步融合”，声音和画面同时发生，大脑就把它们焊成一个事件，所以拟音有效，糟糕的口型同步让人难受。然后是架构：声音和画面有四种接法，先画后声、先声后画、双塔、单一序列。LTX-2 是双塔：140 亿参数的视频流和 50 亿参数的音频流在每个块里互相关注，只用时间位置，两秒处的视频 token 关注两秒处的音频。共享时钟就是全部诀窍。最快的人工检查：看 p、b、m，嘴唇必须闭合。

Revisit: L08 · 6 (Synchresis: why sound is half the picture), L08 · 19 (Inside LTX-2: two streams, one clock), L08 · 41 (Sounds you can see: phonemes → visemes)

48. Reading a lip-sync score, and its trap

Lecture 8 gave you both an eye test and a number for lip-sync. The eye test is phonemes and visemes. Many sounds share one mouth shape, and the bilabials p, b and m need the lips fully closed. If a character says "Bob" and the lips never meet, everyone notices. The number is SyncNet, from Oxford in 2016. It takes a fifth of a second of mouth and audio, embeds both, slides the audio up to fifteen frames each way, and reports Sync-D, the best distance, where lower is better, and Sync-C, the confidence, where higher is better. Then the trap. On LRS2, real video scores about 6.7 Sync-D, and Wav2Lip scores 6.4, better than reality, because it was trained against a SyncNet-style judge. When a metric becomes a training target, it stops measuring what you care about. I promised we'd return to that in Lecture 10, and we did.

第 8 讲给了口型同步一个肉眼测试和一个数字。肉眼测试是音素和视位：很多音共用一个口型，双唇音 p、b、m 必须完全闭唇，角色说“Bob”嘴唇却没合上，所有人都会注意到。数字是牛津 2016 年的 SyncNet：取五分之一秒的嘴部和音频，各自嵌入，把音频前后滑动最多 15 帧，报告 Sync-D（最小距离，越低越好）和 Sync-C（置信度，越高越好）。然后是陷阱：在 LRS2 上，真实视频 Sync-D 约 6.7，Wav2Lip 是 6.4，比现实还好，因为它就是对着类似 SyncNet 的评委训练的。指标一旦成为训练目标，就不再衡量你关心的东西。我答应过第 10 讲会回到这一点，我们也做到了。

Revisit: L08 · 43 (Under the hood: how we measure lip-sync (SyncNet)), L08 · 41 (Sounds you can see: phonemes → visemes), L08 · 44 (Reading the scoreboard (and its trap))

49. Sync starts in the data and in the positions

Two technical reasons joint audio-video models stay in sync. First, positions. Video and audio tokens arrive at different rates. If you number them naively, video token thirty and audio token thirty are at different moments, so

attention looks at the wrong sound. MMAudio's figure showed it: with default positions the bright band drifts, and with positions scaled to real time you get a clean diagonal. LTX-2's temporal-only RoPE and Ovi's scaled RoPE are the same fix. Second, data, which Lecture 8 called the unglamorous truth. Most online video is bad training data for sound: music pasted over vlogs, dubbing, off-screen voice-overs. So teams filter with a sync score and caption the sound, with spoken words in quotation marks. TalkVid, from our lab, went from over 6,000 hours through speech, quality, motion and head filters to 1,244 hours, and the filters agreed with human judges about 95 percent of the time. Good datasets are mostly decisions about what not to keep.

联合音视频模型保持同步的两个技术原因。一是位置：视频和音频 token 到达速率不同，天真地编号，第 30 个视频 token 和第 30 个音频 token 处在不同时刻，注意力就看错了声音。MMAudio 的图显示：默认位置时亮带漂移，按真实时间缩放后是一条干净的对角线。LTX-2 只用时间的 RoPE 和 Ovi 的缩放 RoPE 是同一个修正。二是数据，第 8 讲称之为不光鲜的真相：网上大多数视频对声音来说是坏数据，vlog 配乐、配音、画外音。所以团队用同步分数过滤，并给声音写说明，台词放在引号里。我们实验室的 TalkVid 从 6000 多小时经过语音、质量、运动、头部筛选降到 1,244 小时，筛选结果与人工判断约 95%一致。好的数据集主要是关于不保留什么的决定。

Revisit: L08 · 21 (Keeping time: why positions must be aligned), L08 · 26 (Data: finding clips that are really in sync), L08 · 78 (The real filter funnel, and what it keeps)

50. Directing a performance, not a prompt

Story 2 was about direction. A performance has six parts: voice, timing, mouth, face and eyes, body, and listening, which actors say is half the job. The technical lineage moved from repainting a mouth with Wav2Lip, to a whole head from one photo with VASA-1, to full body and singing with OmniHuman. VASA-1's lesson: generate motion, not pixels, then render. OmniHuman-1.5 added intention: a slow-thinking planner of language-model agents writes emotion, action and shot for each moment, and a fast diffusion transformer acts it out. And then the practical table: an emotion label like "angry" gives you the average angry face from the training data. Directors use playable actions instead. "Angry but held in" becomes "through clenched teeth, without raising his voice." And the strongest option for artists: record the line yourself.

故事二讲导演。表演有六部分：声音、节奏、嘴、脸和眼睛、身体、倾听，演员说倾听占一半。技术谱系：Wav2Lip 重画嘴，VASA-1 从一张照片生成整个头部，OmniHuman 做全身和唱歌。VASA-1 的经验：生成运动而不是像素，再渲染。OmniHuman-1.5 加入意图：由语言模型智能体组成的“慢思考”规划器为每一刻写下情绪、动作、镜头，“快行动”的扩散 Transformer 去演。实用表格：“愤怒”这样的情绪标签只会给你训练数据里平均的愤怒脸；导演用可表演的动作，“压着的愤怒”变成“咬着牙，不提高声音”。对艺术家最有力的选项：自己录台词。

Revisit: L08 · 57 (Directing the delivery: what to ask for), L08 · 39 (What is a performance made of?), L08 · 47 (How VASA-1 works: generate motion, not pixels), L08 · 51 (OmniHuman-1.5: a planner thinks, a renderer acts)

51. Characters that listen: the real-time budget

Story 3 made Lectures 6 and 7 pay off. Real time means two numbers: throughput, meaning forty milliseconds per frame at twenty-five frames per second, and latency, how long after you stop talking the character answers. People leave a gap of about two hundred milliseconds between turns, even though planning a reply takes over six hundred, because we predict the end of the other person's sentence. A normal video diffusion model can't do this: thirty to fifty steps, doubled by guidance, and bidirectional attention, so the first frame isn't finished until the last. The fixes you already knew: chunks with a KV cache, distillation, as in TalkingMachines going from twenty-four evaluations to two, and quantization. Avatar Forcing at CVPR this year added listening: the avatar takes your audio and head motion as input, so it nods and laughs while you speak.

故事三让第 6、7 讲得到回报。实时有两个数字：吞吐，每秒 25 帧即每帧 40 毫秒；延迟，你说完后角色多久回应。人与人轮流说话的间隔约 200 毫秒，尽管计划回答要 600 毫秒以上，因为我们会预测对方句子的结尾。普通视频扩散模型做不到：30 到 50 步，引导再翻倍，双向注意力意味着最后一帧完成前第一帧也不能完成。办法你已经知道：带 KV 缓存的分块、蒸馏（TalkingMachines 从 24 次计算到 2 次）、量化。今年 CVPR 的 Avatar Forcing 加入了倾听：数字人把你的声音和头部动作作为输入，你说话时它会点头、会笑。

Revisit: L08 · 62 (Two numbers that define “real time”), L08 · 64 (Why a normal video diffusion model can’t talk back), L08 · 65 (Three tricks you already know), L08 · 71 (Avatar Forcing: an avatar that listens (CVPR 2026))

52. The real-time toolkit, gathered from four lectures

And here's the toolkit that kept coming back. Change less: adapters like LoRA. Step less: distillation, a student that jumps where the teacher would land. Helios went from about fifty steps to three, and TalkingMachines from twenty-four model evaluations to two. Store less: quantization, Yaofu's Week 6 talk, eight or four bits per number. Remember smartly: generate in chunks, look only backwards, and keep a KV cache so the past isn't recomputed. Why it matters: at twenty-five frames per second you have forty milliseconds per frame, and in conversation people expect a reply in about two hundred milliseconds. Lecture 9 added the payoff: Gaussian splats went from about 0.07 to over a hundred frames per second. Speed changes what something is.

这是反复出现的工具箱。改得更少：LoRA 等适配器。走得更少：蒸馏，学生直接跳到老师会到的地方；Helios 从约 50 步到 3 步，TalkingMachines 从 24 次模型计算到 2 次。存得更少：量化，第 6 周 Yaofu 的讲座，每个数用 8 位或 4 位。聪明地记忆：分块生成、只往回看、保留 KV 缓存，不重算过去。为什么重要：每秒 25 帧，每帧只有 40 毫秒；对话中人们期待约 200 毫秒内回应。第 9 讲的回报：高斯泼溅从约每秒 0.07 帧到 100 帧以上。速度改变了东西的性质。

Revisit: L05 · 67 (Smaller changes, fewer steps, smaller numbers), L07 · 15 (Under the hood: the KV cache), L08 · 65 (Three tricks you already know)

53. Who is the performer?

Story 4. Tilly Norwood, an "actress" who never existed, created by Particle6. The backlash in 2025 was immediate, and this year she was announced as the lead in a feature, Misaligned. Then what the actors won: the 2026 SAG-AFTRA contract, ratified by over ninety-one percent. Producers say they won't use a synthetic performer in a human role without significant additional value, and must go to the union first. The negotiator called it a strong disincentive, not a ban. For digital replicas: informed consent and fair pay, including for actors who have died. And voice: in Lehrman and Sage versus Lovo, voice actors paid a few hundred dollars for "research" recordings found clones of their voices for sale, and a New York court let their claims go forward. Two flashpoints this year had the same tools and opposite receptions: a viral fake fight between two stars without consent, and Dreams of Violets at Tribeca, made with disclosure.

故事四。Tilly Norwood，一个从未存在的“女演员”，由 Particle6 创造。2025 年反对声立刻出现，今年她被宣布担任长片《Misaligned》的主角。演员们争取到的：2026 年 SAG-AFTRA 合同，超过 91%投票通过。制片方承诺除非有显著额外价值，不在人类角色中使用合成表演者，并须先与工会协商，谈判代表称这是强力抑制而非禁令。数字复制品：知情同意和公平报酬，包括已故演员。声音：Lehrman 和 Sage 诉 Lovo 案，配音演员拿几百美元录“研究用”录音，后来发现声音克隆在出售，纽约法院允许诉讼继续。今年两个热点用同样的工具却得到相反的反应：未经同意的两位明星打斗病毒视频，和在翠贝卡首映、做了披露的《Dreams of Violets》。

Revisit: L08 · 85 (Tilly Norwood: an “actress” who never existed), L08 · 88 (What the actors won: the 2026 SAG-AFTRA contract), L08 · 90 (Your voice is part of your identity), L08 · 92 (Two flashpoints from this year)

54. Worlds you can walk into, robots that learn by watching

Lecture 9, October 27th. It opened with a question: I push this cup off the edge of the table, what happens next? A video model can draw the answer. A world model has to answer what happens if you push it. The four stories: from video to playable worlds; 3D is back as Gaussian splats; robots that learn by watching people; and does a world model need pixels at all? We noticed you had already seen two of the slides: Lecture 3 introduced world models, the round-trip test and VLAs, and Lecture 7 gave us the KV cache and the dancing robot.

第 9 讲，10 月 27 日。开场问题：我把这个杯子推下桌子，接下来会怎样？视频模型可以画出答案，世界模型要回答“如果你推它”会怎样。四个故事：从视频到可玩的世界；3D 以高斯泼溅的形式回归；机器人看人学习；世界模型到底需不需要像素？我们注意到你们已经看过其中两页：第 3 讲介绍过世界模型、往返测试和 VLA，第 7 讲给了我们 KV 缓存和跳舞的机器人。

Revisit: L09 · 1 (Worlds You Can Walk Into, Robots That Learn by Watching), L09 · 11 (Genie 3: a world that waits for you), L09 · 50 (λ0 (26 Oct 2026): a humanoid that learned from people)

55. Add an action, and video becomes a place

CMU's definition: a world model takes the previous state and an action and predicts the next state. A video model answers "what happens next?" A world model answers "what happens next if I do this?" Under the hood the network can be almost the same autoregressive video model as Helios, but the action arrives every step, so it can't generate the clip in advance. It has to wait for you. Where do actions come from if the internet has no key-press labels? Genie's latent action model looks at two frames and squeezes the change into one of only eight codes, which turn out to mean left, right and jump. At play time it's a loop: tokenize the frame, turn your key press into a latent action, predict the next tokens, decode, twenty-four times a second. Genie 3 runs at 720p in real time, with promptable world events like "a storm starts".

CMU 的定义：世界模型输入前一状态和一个动作，预测下一状态。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”。底层网络几乎可以和 Helios 一样是自回归视频模型，但动作每一步都到来，所以不能提前生成整段，必须等你。互联网没有按键标注，动作从哪来？Genie 的潜在动作模型看两帧，把变化压缩成仅八个代码之一，结果对应左、右、跳。游玩时是一个循环：帧变 token，按键变潜在动作，预测下一批 token，解码，每秒 24 次。Genie 3 以 720p 实时运行，还能用文字触发世界事件，比如“暴风雨开始”。

Revisit: L09 · 8 (What is a world model?), L09 · 15 (Under the hood: one extra input), L09 · 16 (Under the hood: where do the actions come from?), L09 · 17 (Genie at play time: three parts in a loop)

56. Memory: walk away and come back

The round-trip test started in Lecture 3, with Astra and a doorway. Lecture 9 turned it into the main test of any world model. 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's in its context window. Genie 3 holds visual memory for about a minute. Lecture 3 added an important caution: the test checks spatial and temporal agreement under a short sequence of actions, not all of physics. Contact, materials and control need other tests. And the same capability is now serious work: Waymo built a variant of Genie 3 that also generates lidar, so engineers can rehearse the rare, dangerous situations you can't film safely.

往返测试始于第 3 讲的 Astra 和门口，第 9 讲把它变成任何世界模型的主要测试。看见一扇红门，转身十秒，再转回来：好的世界模型显示同一扇红门，弱的会变成蓝门或者门不见了，因为它只记得上下文窗口里的东西。Genie 3 的视觉记忆约一分钟。第 3 讲补充了重要提醒：这个测试检查的是短动作序列下的空间和时间一致性，不是全部物理，接触、材质和控制需要别的测试。同样的能力现在用于严肃工作：Waymo 做了能同时生成激光雷达数据的 Genie 3 变体，让工程师演练无法安全拍摄的罕见危险情况。

Revisit: L09 · 19 (Under the hood: why the world must come back the same), L03 · 85 (World models: test a round trip), L09 · 13 (Waymo World Model: Genie 3 learns to drive)

57. Drift: one problem in three costumes

This is the slide I most want you to remember tonight. In Lecture 7, long video drifts: the model trains on perfect previous frames but generates from its own imperfect ones. That's the train-test gap, or exposure bias. Self Forcing trains on its own rollouts, and Helios shows the model drifted video on purpose. In Lecture 9, a world model forgets: turn away from the red door, turn back, and it's blue. Genie 3 holds about a minute of visual memory. Also in Lecture 9, a robot copying an expert makes a small error, lands somewhere the expert never went, and makes a bigger one. It's the same shape three times: you train on clean data and then live on your own mistakes. The fixes rhyme too: train on your own outputs, keep a memory, and practise.

今晚我最希望你们记住这一页。第 7 讲：长视频漂移，模型用完美的前一帧训练，却用自己不完美的帧生成，这就是训练-测试差距（暴露偏差）；Self Forcing 用自己的生成结果训练，Helios 故意给模型看漂移的视频。第 9 讲：世界模型会忘记，转身再转回来，红门变蓝了，Genie 3 的视觉记忆约一分钟。还是第 9 讲：模仿专家的机器人犯一个小错，到了专家没去过的地方，犯更大的错。同一个形状出现三次：在干净数据上训练，却生活在自己的错误里。解决办法也押韵：用自己的输出训练、保持记忆、练习。

Revisit: L07 · 21 (Under the hood: the train-test gap), L09 · 19 (Under the hood: why the world must come back the same), L09 · 48 (Under the hood: small mistakes compound)

58. 3D is back: from NeRF to Gaussian splats

Story 2. Pixels or geometry? Genie draws the next frame: alive, but forgetful, with nothing to export. Marble, Atlas and HY-World build a scene first: a file you can open in Blender, Unity or a headset, consistent forever, but mostly static. The geometry side comes from Stanford's two steps. NeRF in 2020 made a scene a function: ask a network what colour and density sit at a point seen from a direction, many times per ray. It's beautiful, but slow. Gaussian splatting in 2023 put millions of soft coloured blobs only where something exists, and threw them onto the screen like paint. Each blob has a position, shape, opacity and colour. Rendering went from about 0.07 frames per second, to about nine with Instant-NGP, to over a hundred. This autumn splats became a normal file format, in Blender, three.js and glTF. Speed changed what it is.

故事二。像素还是几何？Genie 画下一帧：有生命感但健忘，没东西可导出。Marble、Atlas、HY-World 先建场景：一个能在 Blender、Unity 或头显中打开的文件，永远一致，但大多是静态的。几何一侧来自斯坦福讲的两步：2020 年的 NeRF 把场景变成函数，问网络某点从某方向看是什么颜色和密度，每条光线问很多次，很美但很慢；2023 年的高斯泼溅只在有东西的地方放数百万个柔和的彩色小球，像泼颜料一样投到屏幕上，每个球有位置、形状、不透明度和颜色。渲染从约每秒 0.07 帧，到 Instant-NGP 的约 9 帧，再到 100 帧以上。今年秋天泼溅成了常规文件格式，进入 Blender、three.js 和 glTF。速度改变了它的性质。

Revisit: L09 · 37 (The result: from 0.07 to 100+ frames per second), L09 · 27 (Pixels or geometry? Two ways to make a world), L09 · 33 (Under the hood: from a scene to a function (NeRF)), L09 · 34 (Under the hood: only put stuff where stuff is), L09 · 40 (This autumn, splats became a normal file format)

59. Generate the small thing, render the big thing

Here's a pattern that showed up in three lectures without being named. VASA-1 doesn't generate pixels for a talking face. It generates a tiny motion code for head pose and expression, then renders frames from your photo, which is why it runs at forty frames per second. Genie learned actions nobody labelled by squeezing the change between two frames into one of only eight codes, which turned out to mean left, right and jump. Gaussian splats put soft blobs only where something exists instead of asking a network about empty air. In each case the creative and technical win came from choosing what to represent. When your own project is too slow or unstable, ask what the small thing is.

这个模式在三讲中出现过，但没有命名。VASA-1 不为说话的脸生成像素，它生成一个很小的运动编码（头部姿态和表情），再从你的照片渲染帧，所以能达到每秒 40 帧。Genie 把两帧之间的变化压缩到仅八个代码之一，学会了没人标注的动作，结果对应左、右、跳。高斯泼溅只在有东西的地方放柔和的小球，而不是向网络询问空气。每次的创意和技术突破都来自选择表示什么。你的项目太慢或不稳定时，问问：那个“小东西”是什么？

Revisit: L08 · 47 (How VASA-1 works: generate motion, not pixels), L09 · 16 (Under the hood: where do the actions come from?), L09 · 34 (Under the hood: only put stuff where stuff is)

60. Robots that learn by watching people

Story 3 brought Lectures 3 and 7 together. A VLA starts from a vision-language model and attaches an action expert that uses diffusion or flow matching, the same generation method from Lecture 3, to output joint commands instead of pixels, about a second of actions at a time. The data problem is that robot demonstrations are scarce. λ0 from Amap solved it with HumanVerse: people wear a camera at the robot's head height, plus body and hand trackers, and just live, which is about 3.4 times faster than collecting robot data. Then three stages: learn hand-object interaction from human video, translate it into the robot's body, and finish with six hundred real demos. It reached 62.5 percent on four tasks against forty for π0.5, and only fifteen without human data. Copying alone plateaus around eighty percent, and homes need ninety-nine, so practice through reinforcement learning, often inside world models, is the next step.

故事三把第 3 讲和第 7 讲合在一起。VLA 从视觉语言模型出发，接一个动作专家，用第 3 讲的扩散或流匹配输出关节指令而不是像素，每次约一秒的动作。数据问题是机器人示范稀缺。高德的 λ0 用 HumanVerse 解决：人在机器人头部高度佩戴相机和身体、手部追踪器，正常生活，比收集机器人数据快约 3.4 倍。三个阶段：从人类视频学手物交互，转换到机器人身体，最后用 600 条真实示范收尾。四项任务成功率 62.5%，π0.5 是 40%，去掉人类数据只有 15%。单靠模仿在约 80%停滞，家庭需要 99%，所以下一步是强化学习式的练习，常常在世界模型里进行。

Revisit: L09 · 53 (λ0, by the numbers), L09 · 46 (Under the hood: what is inside a VLA?), L09 · 51 (HumanVerse: wear the robot’s eyes), L09 · 59 (Under the hood: why copying plateaus around 80%)

61. Does a world model need to draw pixels?

Story 4 was a debate between two camps. One generates the world: predict the next frame in pixels or splats, so you can watch the prediction, train robots inside it and sell it as video. That's Genie, Gemini Omni, Atlas and NVIDIA's Cosmos 3, an open model with a reasoner and a generator that outputs video and robot actions. The other camp predicts in latent space: hide part of a video and predict an abstract summary of the missing part, never its pixels. That's Yann LeCun's JEPA. V-JEPA 2 was pretrained on over a million hours of video, then adapted with under sixty-two hours of robot video to plan moves. It imagines action sequences in embedding space, picks the one closest to the goal image, takes one step and re-plans, without drawing a single frame. The motion was: a video generator is a world model. We said robots will decide.

故事四是两个阵营的辩论。一方生成世界：用像素或泼溅预测下一帧，你能看到预测、在里面训练机器人、还能当视频卖，例如 Genie、Gemini Omni、Atlas 和 NVIDIA 的 Cosmos 3（开源，含推理器和能输出视频与机器人动作的生成器）。另一方在潜空间预测：遮住视频的一部分，预测缺失部分的抽象摘要，从不画像素，这是杨立昆的 JEPA。V-JEPA 2 用超过一百万小时视频预训练，再用不到 62 小时机器人视频适配来规划动作：在嵌入空间里想象动作序列，选最接近目标图的那个，走一步再重新规划，一帧都不画。辩题：视频生成器就是世界模型。我们说：机器人会做出裁决。

Revisit: L09 · 65 (Two camps), L09 · 66 (NVIDIA Cosmos 3: one open model to reason, simulate and act), L09 · 68 (Why not pixels? JEPA predicts the meaning, not the picture), L09 · 69 (V-JEPA 2-AC: plan a robot move without drawing the future)

62. Is it any good?

Lecture 10, last week, November 3rd. It began with a game: score a picture from one to ten, learn that it's Théâtre D'opéra Spatial, the AI image that won the Colorado State Fair and was refused copyright, and score it again. The picture didn't change by one pixel; only what you knew about it changed. The four stories: looking, measuring, showing and signing, about two thirds art and one third technique. It pulled together questions you'd been circling all term, and it's the lecture closest to what you'll do in your final presentation, because there you have to argue that your work is good and that it's honestly yours.

第 10 讲，上周，11 月 3 日。以一个游戏开场：给一张图打分，得知它是在科罗拉多州博览会获奖、又被拒绝版权登记的 AI 图像《太空歌剧院》，再打一次分。图一个像素都没变，变的只是你知道的信息。四个故事：看、量、展、署名，约三分之二艺术、三分之一技术。它把整个学期绕着的问题收拢起来，也是最接近期末展示的一讲，因为在那里你要论证你的作品是好的，而且诚实地属于你。

Revisit: L10 · 1 (Is It Any Good? Looking at, measuring, showing and signing AI art), L10 · 3 (Now you know: it won a prize, then it was refused copyright), L10 · 80 (Generating is now cheap. Looking, deciding, and signing your name are not.)

63. Every new medium gets asked: is this even art?

Story 1 started with history, because the question is old. In 1859 Baudelaire called photography the refuge of every failed painter, and then he sat for Nadar. In 1917 Duchamp's friends defended a signed urinal with one sentence: "He CHOSE it." That's still the best argument for AI art, with the warning that Duchamp's choice cost him something. Around 1969 Vera Molnár drew with a plotter and was mostly ignored. Then this year: on October 2nd Pope Leo told artists there's an ontological difference between art and what a machine generates, and three weeks later an artist answered him in an open letter. And Merriam-Webster's word of 2025 was "slop": low quality, in quantity. What's the opposite of slop? Not handmade. I've seen plenty of handmade rubbish. The opposite of slop is decided.

故事一从历史开始，因为这个问题很古老。1859 年波德莱尔说摄影是失败画家的避难所，后来他自己让纳达尔拍了照。1917 年杜尚的朋友用一句话为签了名的小便池辩护：“他选择了它。”这仍是为 AI 艺术辩护的最好论点，但要记得杜尚的选择是有代价的。约 1969 年，Vera Molnár 用绘图仪作画，几乎被忽视。今年 10 月 2 日，教皇良十四世对艺术家说艺术和机器生成之间有本体论差异，三周后一位艺术家发表公开信回应。韦氏词典 2025 年度词是“slop”：低质量、大批量。slop 的反义词是什么？不是手工，我见过太多手工垃圾。slop 的反义词是“经过决定的”。

Revisit: L10 · 12 (Slop: the word of 2025, and what its opposite is), L10 · 9 (1859: Baudelaire hated photography. Then he sat for Nadar.), L10 · 10 (1917: the first prompt engineer? “He CHOSE it.”), L10 · 11 (A month ago: the Pope and an artist, in public)

64. Four questions people mix up

Last week's most useful slide, with the whole semester mapped onto it. Is it AI? That's fact, answered with evidence: credentials, watermarks, the maker's log. Is it good? Judgment, answered by looking and arguing. No machine settles it, and Lecture 4 already warned that a score answers a narrow question. Is it fair? Ethics: whose data, whose consent, whose likeness, whose job. That's Lectures 5, 8 and 9. Is it allowed? Law, and it differs by country and by festival: the lawsuits in Lecture 5, the SAG-AFTRA contract in Lecture 8, the EU's Article 50 last week. Most fights online happen because someone asks question two and someone else answers question three. Remember the label experiment from that night: telling people an image was made by a human barely changed how much they liked it, but changed how much they thought it was worth by more than three times as much. In your final presentation, say which question you're answering.

上周最有用的一页，加上整个学期的对应。是 AI 吗？事实问题，用证据回答：凭证、水印、创作者的记录。好吗？判断问题，靠观看和论证，没有机器能裁定，第 4 讲已提醒分数只回答狭窄的问题。公平吗？伦理问题：谁的数据、谁的同意、谁的肖像、谁的工作，即第 5、8、9 讲。允许吗？法律问题，因国家和电影节而异：第 5 讲的诉讼、第 8 讲的 SAG-AFTRA 合同、上周的欧盟第 50 条。网上大多数争吵是因为有人问第二个问题，别人回答了第三个。期末展示时，请说明你在回答哪个问题。

Revisit: L10 · 13 (The most useful slide tonight: four questions people mix up), L04 · 44 (A technical score answers

a narrow question)

65. Look before you judge

So if scores can't decide, how do you judge? With methods. See, Think, Wonder from Harvard's Project Zero: what do you see, only facts; what do you think about it; what does it make you wonder? Liking comes last. Feldman's four steps: describe, analyze, interpret, and only then judge. We did it on Edmond de Belamy. And the four questions from Eryk Salvaggio's reading, made for AI images: what's here too often, what's missing, where does it break, and most important for you as makers, what did the artist fight? Which part goes against the default? That's where your authorship lives, and it's what the panel will look for on the 17th and 24th.

如果分数不能决定，怎么判断？用方法。哈佛 Project Zero 的“看-想-疑”：看到什么，只说事实；你怎么想；它让你好奇什么？喜欢放在最后。Feldman 四步：描述、分析、解释，最后才评判，我们用《Edmond de Belamy》练过。Eryk Salvaggio 读物中为 AI 图像设计的四个问题：什么出现得太多？缺了什么？哪里崩坏？对创作者最重要的是：艺术家对抗了什么？哪部分违背了默认？那就是你的作者身份所在，也是 17 日和 24 日评审要看的。

Revisit: L10 · 17 (See. Think. Wonder.), L10 · 18 (A critic’s four steps, on a work you know: Edmond de Belamy), L10 · 20 (How to read an AI image: four questions to add to your crit)

66. Measuring: what a number can see

Story 2 was the technical part, adapted from CMU and Stanford. We measured it ourselves that afternoon. PSNR compares pixel to pixel, so a picture moved by eight pixels, which you can't tell from the original, scores worse than a wrecked JPEG. FID passes thousands of images through a network, fits a blob to each cloud of points, and measures the distance. It judges a crowd, never one image, so there's no such thing as the FID of your film, and a copycat that replays its training set scores a perfect zero. VBench stops giving one number and gives sixteen, like a critic saying "strong motion, weak hands." Elo from blind votes: a hundred points means winning about 64 percent of votes, and forty-three points is about 56, nearly a coin flip. And the LAION aesthetic score, trained on ratings from a photo contest site, chose Stable Diffusion's training data. Vera Molnár scored 4.87, below the cut-off.

故事二是技术部分，改编自 CMU 和斯坦福，那天下午我们亲自测了。PSNR 逐像素比较，平移 8 像素、肉眼无法分辨的图，得分比严重损坏的 JPEG 还差。FID 把成千上万张图过网络，给每团点拟合一个团，量两团之间的距离；它评判一群图而非一张，所以不存在“你的电影的 FID”，只复述训练集的抄袭者得到完美的 0。VBench 不给一个数而给十六个，像评论家说“动作强、手弱”。盲投的 Elo：差 100 分约赢 64%的票，差 43 分约 56%，几乎是抛硬币。LAION 美学分用摄影比赛网站的评分训练，选择了 Stable Diffusion 的训练数据；Vera Molnár 得 4.87，低于门槛。

Revisit: L10 · 31 (Worked example: the copycat gets a perfect score), L10 · 28 (What pixels say: moving 8 pixels is worse than wrecking it), L10 · 34 (VBench: stop giving one number. Give sixteen.), L10 · 40 (We ran that predictor

on art history this afternoon)

67. Every score we met, and its trap

Here's every score we met this term on one slide, each with its trap. Raw frame difference, Lecture 2: the frozen video wins. PSNR and SSIM, Lecture 10: an image shifted by eight pixels, which you can't tell from the original, scores worse than a wrecked JPEG. FID: it judges a crowd, never one image, and a copycat that replays its training set scores a perfect zero. Sync-C and Sync-D, Lecture 8: Wav2Lip scored better than real video, because it was trained against that judge. The LAION aesthetic score: Vera Molnár's plotter drawing fell below the training cut-off. Elo: a 43-point gap between first and tenth is about 56 percent of votes, nearly a coin flip. The common law is Goodhart's: when a measure becomes a target, it stops being a good measure. So use scores to compare versions of your own pipeline, and never let one pick your work for you.

本学期遇到的所有分数和它们的陷阱。原始帧差（第 2 讲）：静止视频获胜。PSNR 和 SSIM（第 10 讲）：平移 8 像素、肉眼无法分辨的图，得分比严重压缩的 JPEG 还差。FID：评判一群图而不是一张，只复述训练集的抄袭者得到完美的 0。Sync-C/Sync-D（第 8 讲）：Wav2Lip 分数超过真实视频，因为它就是对着这个评委训练的。LAION 美学分：Vera Molnár 的绘图仪作品低于训练门槛。Elo：第一和第十相差 43 分，约 56%的票，几乎是抛硬币。共同规律是古德哈特定律：指标一旦成为目标，就不再是好指标。用分数比较你自己流程的不同版本，但别让它替你挑作品。

Revisit: L02 · 94 (The frozen video wins the wrong test), L08 · 44 (Reading the scoreboard (and its trap)), L10 · 31 (Worked example: the copycat gets a perfect score)

68. Showing: what happens when the work leaves your laptop

Story 3 went to museums. Dataland opened in June in Los Angeles: 25,000 square feet, Refik Anadol's Large Nature Model, wristbands, scent, even taste. The critics disagreed, and the useful exercise was asking which of our four questions each critic was answering. Museums make claims too, like permission-based data or one iPhone charge per visit, and a good critic asks how you'd check them. Then Madison: Beeple's Diffuse Control transformed the museum's own collection through a diffusion model. A letter signed by more than nine hundred people objected. We found two layers of consent: the collection, which the museum could answer for, and the model's training data, which no single museum can. And at the V&A, Vera Molnár. The same word, "generative", covers very different decisions: she wrote the rule, chose about one percent randomness and could explain every line.

故事三走进美术馆。Dataland 6 月在洛杉矶开馆：25000 平方英尺，Refik Anadol 的大自然模型，手环、气味甚至味道。评论家意见不一，有用的练习是问每位评论家在回答四个问题中的哪一个。美术馆也会提出主张，比如数据经过许可、每次参观耗电相当于给 iPhone 充一次电，好的评论家会问怎么核实。然后是麦迪逊：Beeple 的《Diffuse Control》用扩散模型转化博物馆自己的藏品，900 多人签名的公开信表示反对。我们发现同意有两层：藏品层，博物馆能回答；模型训练数据层，没有哪家博物馆能回答。还有 V&A 的 Vera Molnár：同一个词“生成式”涵盖截然不同的决定，她写规则，选择约百分之一的随机，能解释每一条

线。

Revisit: L10 · 53 (Two layers of consent, and the museum answered one), L10 · 48 (Dataland: “the world’s first museum of AI arts”), L10 · 51 (Madison, 18 September: a museum, Beeple, and a curtain), L10 · 55 (Same word, “generative”. Different decisions.)

69. Consent has layers

Here's the fairness question, gathered from four lectures. Consent has layers, and one answer rarely covers all of them. The training data: Lecture 5's lawsuits and memorized images, and the second layer at Madison, the billions of images behind the model, which no single museum can answer for. Likeness: Lecture 8's Tilly Norwood, and the 2026 SAG-AFTRA contract requiring informed consent and fair pay for digital replicas, including for actors who have died. Places and bodies: Lecture 9 asked who owns the scan of a place, and whose homes become robot training data. And the collection: Madison could answer this layer, because it had the rights to the works on screen. When you write the consent line in your credits, say which layers you can answer.

公平问题，汇总自四讲。同意有很多层，一个回答很少能覆盖全部。训练数据：第 5 讲的诉讼和被记忆的图像，以及麦迪逊案中的第二层，模型背后的数十亿张图，没有哪家博物馆能替它回答。肖像：第 8 讲的 Tilly Norwood，2026 年 SAG-AFTRA 合同要求数字复制品须知情同意并公平付费，包括已故演员。地方与身体：第 9 讲问谁拥有一个地方的扫描、谁的家成为机器人训练数据。藏品：麦迪逊能回答这一层，因为它拥有屏幕上作品的权利。在署名中写同意说明时，说清楚你能回答哪几层。

Revisit: L05 · 16 (Ongoing legal battles (1)), L08 · 88 (What the actors won: the 2026 SAG-AFTRA contract), L10 · 53 (Two layers of consent, and the museum answered one)

70. Sign your work, and don't apologise

And finally, signing. There are three layers with three different promises. A visible label promises only that someone told you. Signed metadata, C2PA, hashes the file, writes a claim, signs it and embeds it. It's tamper-evident, because our edited file read as invalid, but stripping it takes one line of code: a re-save, a screenshot, a resize. So a credential proves a lot when it's there, and its absence proves nothing. The pixel watermark is the third layer. The EU's art exception says you must still disclose, but in a way that doesn't spoil the work, so end credits can be enough. Then use the credit block. It counts your human work: forty-one attempts, six selected, edit and grade. It should read like a statement of what you did, not an apology.

最后是署名。三层，三种不同的承诺。可见标签只承诺有人告诉了你。签名元数据 C2PA：哈希文件、写声明、签名、嵌入；它能显示篡改（我们改过的文件显示无效），但去掉它只需一行代码：重新保存、截图、缩放。所以凭证在时能证明很多，不在时什么也证明不了。像素水印是第三层。欧盟的艺术例外：仍须披露，但方式不应妨碍作品，片尾字幕可能就够了。然后用署名块，数出你的人工工作：41 次尝试、选 6 个、剪辑调色。它读起来应该像作品说明，而不是道歉。

Revisit: L10 · 64 (Under the hood: three layers, three different promises), L10 · 67 (We broke it four ways, in Python, in five minutes), L10 · 73 (Useful: a credit block you can copy for A2 and A3)

71. Signing under the hood: hash, sign, and a hidden message

The technical half of Story 4, as we tested it that afternoon. C2PA works like the padlock in your browser. Hash: boil the file's bytes down to a fingerprint, and one changed bit gives a completely different fingerprint. Claim: write the facts, such as created, digital source type "trained algorithmic media", and training and mining not allowed. Sign it with a private key, and embed it in the file. Note what it doesn't say: it doesn't say the claim is true, only that this signer said this about this exact file. Our demo also read "signing credential untrusted", because nobody had agreed to trust my certificate. Then we broke it: a JPEG re-save, a screenshot and a resize all removed it silently, while editing pixels with the manifest kept read as invalid, a hash mismatch. Last, the watermark. The open DWT-DCT method kept a hundred percent of its bits untouched, sixty-seven percent after a normal JPEG save, against fifty for a coin flip, and almost nothing after a crop, rotation or blur. SynthID is a learned watermark built to survive that, and it's what caught the Nikon video.

故事四的技术部分，按那天下午的测试。C2PA 像浏览器里的小锁：哈希，把文件字节压成指纹，改一位指纹就完全不同；声明，写下事实，比如“创建”、数字来源类型“训练过的算法媒体”、不允许训练和挖掘；用私钥签名；嵌入文件。注意它没说的：它不说声明为真，只说这个签名者对这个确切文件说了这些。我们的演示还显示“签名凭证不受信任”，因为没人同意信任我的证书。然后我们攻破它：重存 JPEG、截图、缩放都悄无声息地去掉了凭证；保留清单修改像素则显示无效，哈希不匹配。最后是水印：开源的 DWT-DCT 方法未改动时保留 100%比特，普通 JPEG 保存后 67%，而抛硬币是 50%，裁剪、旋转、模糊后几乎全无。SynthID 是学习式水印，专门为抵抗这些而设计，也正是它识破了尼康的视频。

Revisit: L10 · 65 (Under the hood: C2PA in four steps), L10 · 67 (We broke it four ways, in Python, in five minutes), L10 · 68 (Under the hood: the watermark. A hidden message, tested.), L10 · 66 (See it: the real manifest we signed this afternoon)

72. Bring the semester into your final presentation

Now let's turn the review into your final presentation. The course promised six outcomes, and each one maps to something from this term. Explain a pipeline: draw your own four-stage diagram with real model names and versions. Direct a coherent work: show decisions, not just outputs, with your edit contract or beat sheet. Build and document an experiment: one variable at a time, attempts counted, with a side-by-side, as Lecture 1 asked. Critique with evidence: run the four steps on your own work and say what you fought against the default. Account for AI use: your credit block, with versions pinned and a consent line. Present a refined project: one clear intention, and what you'd do next. If each of these gets one slide, you have a strong presentation.

把回顾变成你的期末展示。课程承诺六个学习成果，每个都对应本学期的内容。解释流程：画出你自己的四阶段图，写上真实模型名和版本。导演一个连贯作品：展示决定，而不只是输出，用你的编辑契约或节拍表。构建并记录实验：一次一个变量，数清尝试次数，配并排对比，正如第 1 讲要求的。用证据评论：

对自己的作品走四步，说出你对抗了哪些默认。说明 AI 的使用：署名块、锁定版本、同意说明。呈现一个完善的项目：一个清晰的意图，以及下一步做什么。每项一页，就是一场很好的展示。

Revisit: L01 · 101 (Design one eight-second shot before opening a generator), L10 · 81 (Crit sheet: copy this for your final presentation), L10 · 73 (Useful: a credit block you can copy for A2 and A3)

73. Three weeks left

The practical part. Tonight, after the break, is the last week for weekly presentations. November 17th and 24th are the final presentations, so check your slot on the sign-up page and add your slides link. Assignments 2 and 3 are both due November 30th at 23:59 Hong Kong time. Every quiz, including tonight's Quiz 11, is due at the end of the semester, and you can retrieve and edit your answers until then. Quiz 11 is a review quiz, and each question asks you to connect two lectures, so it doubles as preparation for your final.

实际安排。今晚休息后是最后一周的每周展示。11 月 17 日和 24 日是期末展示，请在报名页确认时段并添加幻灯片链接。作业二和作业三都在 11 月 30 日香港时间 23:59 截止。所有小测验，包括今晚的第 11 次，都在学期末截止，在那之前可以随时取回修改。第 11 次是复习测验，每题都要求连接两讲内容，也可以当作期末准备。

74. Technical literacy becomes creative agency

Let me end with two sentences, one from the first night and one from the last. From Lecture 1: when an output surprises you, don't only ask whether you like it. Ask which decision produced it, and what controlled comparison would teach you more. That's where technical literacy becomes creative agency. And from Lecture 10: generating is now cheap; looking, deciding and signing your name are not. Between those two sentences is everything we did this term. Thank you. Let's take a ten-minute break, and then the last weekly presentations.

最后两句话，一句来自第一晚，一句来自上周。第 1 讲：当输出让你意外时，不要只问喜不喜欢，要问是哪个决定产生了它，什么样的对照比较能让你学到更多。这就是技术素养变成创作能力的地方。第 10 讲：生成如今很便宜，观看、决定、签上你的名字却不便宜。这两句话之间，就是我们这学期做的一切。谢谢大家。休息十分钟，然后是最后一轮每周展示。

Revisit: L01 · 102 (Technical literacy becomes creative agency), L10 · 80 (Generating is now cheap. Looking, deciding, and signing your name are not.)