

AMCC 5160 · Lecture 08 — When AI Video Learned to Talk: Sound, Performance, and Real-Time Characters

Study transcript · English narration with Chinese companion explanations (中文讲解). 20 October 2026.

1. When AI Video Learned to Talk

Good evening, everyone, and welcome to Week 8. For seven weeks, almost everything we generated was silent. Even the beautiful clips in the last lecture, the minute-long Helios videos and the Wan-Dancer dances, needed someone to add the sound afterwards. Tonight is about the moment that changed. Generated video now speaks, sings, breathes and makes its own footsteps, in one pass. Characters can now hold a conversation with you in real time. And that raises an old question in a new form: who is the performer? Tonight has more videos than any lecture so far, so please make sure you can hear the speakers.

中文讲解：欢迎来到第八讲。前七周我们生成的几乎都是无声视频，连上周的Helios长视频和Wan-Dancer舞蹈也需要事后配音。今晚讲的是这个变化：生成视频现在可以一次性生成说话、唱歌、呼吸和脚步声；角色可以实时与你对话。由此也带来一个老问题的新形式：谁是表演者？今晚的视频比以往任何一讲都多，请确保能听到声音。

2. Sound off: what are they saying?

Let's start with a game. I'm going to play a five-second clip with the sound off. [Play muted.] Two characters, a dark room, a robot. Turn to your neighbour for twenty seconds: what is being said? What do the voices sound like? What else would you hear in that room? Keep your guesses. Notice that you are already doing what film audiences always do: your brain is filling in sound from the image.

中文讲解：先玩个游戏：静音播放一段5秒片段。两个角色、昏暗房间、一个机器人。和邻座讨论20秒：他们在说什么？声音是什么样？房间里还会有什么声音？记住你的猜测。你的大脑已经在根据画面“补”声音了。

3. Sound on: one model made all of it

Now with the sound on. [Play with sound.] How close were you? Here is the point for tonight: nobody recorded these voices, and nobody synced them. The picture, the two voices, the reverb of the room and the timing of every word came out of one model, in one generation. This clip is from Ovi, an open model released by Character.AI in 2025. A year and a half ago, making this would have taken four or five separate tools and an editor.

中文讲解：现在打开声音再看一次。你猜对了多少？今晚的重点：没有人录这些声音，也没有人做口型对齐。画面、两个声音、房间混响和每个字的时机，都由一个模型一次生成。这段来自Character.AI在2025年开源的Ovi。一年半前，这需要四五个工具外加剪辑师。

4. Four stories from this year

Here is the map for tonight. Story one: sound is no longer post-production. Since Google's Veo 3 in May 2025, the leading video models generate picture and sound together, and this month the open models caught up. Story two: lip-sync grew into acting. The question is no longer 'do the lips match', but 'is this a good performance, and how do I direct it'. Story three: characters that listen. Real-time avatars now react while you are still speaking. Story four: who is the performer? An AI 'actress' has been cast in a feature film, and the actors' union has written new rules. We'll have presentations and a break after story two, and we'll finish with a debate.

中文讲解：今晚四个故事：一、声音不再是后期——自2025年5月Veo 3以来，主流模型同时生成画面与声音，本月开源模型也赶上了；二、从口型同步到表演——问题不再是嘴型对不对，而是表演好不好、如何导演；三、会倾听的角色——实时数字人能在你说话时就做出反应；四、谁是表演者——AI“演员”被选角主演电影，演员工会制定了新规则。第二个故事后是学生展示和休息，最后是辩论。

5. Where we've been: seven weeks in one slide

A quick recap, because tonight builds on almost everything. Week 3 taught us that models see tokens, not pixels; tonight sound becomes tokens too. Week 6, Yaofu's quantization talk, and Week 7's Self Forcing and KV cache are exactly the tricks that make real-time characters possible, so story three is where those lectures pay off. And Week 5's question, who owns a model you adapt, becomes much sharper when the thing being copied is a person's voice and face.

中文讲解：快速回顾：第三周讲模型看到的是token而非像素，今晚声音也变成token；第六周耀夫讲的量化和第七周的Self Forcing与KV缓存，正是实时角色的关键技术；第五周“谁拥有你微调的模型”的问题，当被复制的是一个人的声音和面孔时会更加尖锐。

6. Synchresis: why sound is half the picture

Before the technology, one idea from film theory, especially for the art students. The French composer and theorist Michel Chion, in his book Audio-Vision from 1990, coined the word 'synchresis', from synchronism and synthesis. When a sound and an image happen at the same instant, your brain welds them into one event, even if they have nothing to do with each other. That is why Foley works: a sound designer snaps celery for a breaking bone, and you believe it. It is also why bad AI lip-sync is so painful: the welding is automatic, so when the timing is off, everyone notices. Broadcasters measured this: in the ITU's tests, viewers start to notice when sound arrives about 45 milliseconds early or 125 milliseconds late. At 24 frames per second, that is one to three frames. Synchronization is not a detail. It is the whole illusion.

中文讲解：先讲一个电影理论概念。法国作曲家与理论家米歇尔·希翁在1990年的《视听》一书中提出“合成同步”（synchresis）：当声音与画面同时发生，大脑会把它们焊接成一个事件，即使它们毫无关系。这就是拟音有效的原因——折断芹菜可以当作骨折声。也正因如此，AI口型不同步会非常刺眼。国际电信联盟的测试显示，声音提前约45毫秒或滞后约125毫秒，观众就会察觉，也就是24帧下的一到三帧。同步不是细节，而是整个幻觉的核心。

7. Sound is no longer post-production

Story one. For a hundred years of cinema, sound was something you added after the picture: dialogue recorded or re-recorded, Foley performed, music composed and mixed. AI video copied that workflow at first. This story is about how that changed in about eighteen months, and what is happening inside the model when it does.

中文讲解：第一个故事。电影百年来，声音都是在画面之后添加的：对白录制或补录、拟音表演、配乐与混音。AI视频起初也照搬了这个流程。这个故事讲的是它如何在约18个月内改变，以及模型内部发生了什么。

8. The old pipeline: five tools, five chances to drift

Here is how you would have made a talking scene a year ago, and honestly how many of you made Assignment 1. Generate a silent clip. Generate or record the voice. Run a lip-sync model like Wav2Lip, from 2020, which repaints the mouth to match the audio. Run a video-to-audio model like MMAudio to add footsteps and room tone. Then mix it all in an editor. Five tools, and every hand-off is a chance for something to drift: the voice is calm but the face is angry, the footsteps land a frame late, the mouth moves but the cheeks don't.

中文讲解：一年前做一个说话场景的流程（也是很多同学做作业一的方式）：先生成无声视频，再生成或录制配音，用Wav2Lip（2020）之类的模型重绘嘴型，用MMAudio之类的视频转音频模型加脚步声和环境声，最后在剪辑软件里混音。五个工具，每次交接都可能出错：声音平静但脸很愤怒、脚步声晚一帧、嘴在动但脸颊不动。

9. Eighteen months, from add-on to native

The timeline. Lip-sync as a repair step goes back to Wav2Lip in 2020. At the end of 2024, MMAudio made video-to-audio Foley fast, about a second for an eight-second clip. Then in May 2025 Google released Veo 3, the first major model to generate picture and sound jointly, and its demo clips went viral within days. Within a few months the open-source world answered with Ovi from Character.AI, then LTX-2 from Lightricks with open weights in January this year, and ByteDance's Seedance 2.0 in February. In the last three months: MiniMax H3, Kandinsky 6.0, Kling 4.0 and Vidu Q4. Native sound is now the default, not the feature.

中文讲解：时间线：2020年Wav2Lip把口型同步做成独立修补步骤；2024年底MMAudio让视频转音频拟音只需约1秒；2025年5月谷歌Veo 3首次联合生成画面与声音，演示片段几天内爆红；随后Character.AI开源Ovi，今年1月Lightricks

开放LTX-2权重，2月字节发布Seedance 2.0；最近三个月又有MiniMax H3、Kandinsky 6.0、可灵4.0和Vidu Q4。原生声音已成为默认配置。

10. Veo 3: the clip that started it

This is one of the clips Google showed when it launched Veo 3 in May 2025. [Play.] An old sailor, a line of dialogue, wind, the sea. None of it was recorded, and none of it was synced by hand. Google's model card describes Veo 3 as a latent diffusion model that denoises audio latents and video latents together. That one sentence is the whole of story one.

中文讲解：这是谷歌2025年5月发布Veo 3时展示的片段之一：老水手、一句台词、风声、海浪声。没有任何录音，也没有手工对齐。谷歌的模型卡把Veo 3描述为同时对音频潜变量和视频潜变量去噪的潜扩散模型。这一句话就是第一个故事的全部。

11. Three kinds of sound, one model

Three short Veo 3 clips, three kinds of sound. [Play each.] Effects: every toss of the wok has its sizzle and clatter. Music: the bow and the note move together. Dialogue: a detective interrogating a rubber duck, which only works if the comic timing works. Vote with a show of hands: which sound welds to the picture best? Usually it's the effects, because they are tied to one visible event; dialogue is the hardest.

中文讲解：三段Veo 3短片，三种声音：音效——锅每次翻动都有滋滋声和碰撞声；音乐——琴弓与音符同步；对白——侦探审问橡皮鸭，喜剧靠节奏。举手投票：哪种声音与画面“焊接”得最好？通常是音效，因为它绑定在一个可见事件上；对白最难。

12. Who ships native audio in 2026

Here is who ships native audio now. On the closed side: Google's Veo 3.1, ByteDance's Seedance, Kuaishou's Kling 4.0 and ShengShu's Vidu Q4, which came out two weeks ago with voice references. On the open side, and this is the big change this year, you can download and run LTX-2 from Lightricks, Ovi from Character.AI, MiniMax H3, and Kandinsky 6.0, released this month under the MIT license. For your assignments, this means you can now get sound and picture together without paying for an API, if you have a good GPU.

中文讲解：目前支持原生声音的模型：闭源有谷歌Veo 3.1、字节Seedance、快手可灵4.0和生数Vidu Q4（两周前发布，支持声音参考）；开源有Lightricks的LTX-2、Character.AI的Ovi、MiniMax H3和本月以MIT协议发布的Kandinsky 6.0。对你们的作业而言，只要有好显卡，就能免费同时生成声音和画面。

13. Open weights, this summer and autumn

Three open models from the last six months. [Play each.] Kandinsky 6.0, released this month under MIT, here a dog in a living room with room sound. MiniMax H3, a 33-billion-parameter single-stream model, with stereo sound. And Baidu's NAVA, only 6.3 billion parameters, with two chess players talking; it currently has the best lip-sync score among open models. You can download all three. Do check each license before using them commercially; they differ.

中文讲解：过去半年的三个开源模型：本月以MIT协议发布的Kandinsky 6.0（客厅里的狗和室内声）；330亿参数单流的MiniMax H3（立体声）；百度仅63亿参数的NAVA（两位下棋者对话），目前开源模型中口型同步得分最高。三个都可以下载，商用前请分别核对许可证。

14. Closed models: Seedance 2.0's range

And from the closed side, two clips from ByteDance's official Seedance 2.0 page. [Play both.] On the left, a Chinese ink painting comes alive, and the hooves and wind are generated to match the style. On the right, a concert pianist: hands, music, and the reverb of the hall together. Seedance accepts up to nine images, three videos and three audio clips as references, which is why it's popular with commercial studios.

中文讲解：闭源方面，字节Seedance 2.0官方页面的两段：左边一幅水墨画活了过来，马蹄声和风声与画风相配；右边是音乐厅钢琴家：手、音乐和大厅混响一起生成。Seedance最多接受9张图、3段视频和3段音频作为参考，因此在商业工作室很受欢迎。

15. Under the hood: what is sound to a model?

Now under the hood, and I'll keep it visual. This sequence is adapted from Carnegie Mellon's Generative AI course, lecture 24 on audio. Raw sound is a pressure wave, recorded as tens of thousands of numbers per second: forty-eight thousand for studio audio. Nobody feeds that straight into a transformer. First, we run many short Fourier transforms over overlapping windows, which turns sound into an image: time on one axis, pitch on the other. That's a spectrogram, and the 'mel' version warps the pitch axis to match human hearing. Then, just as a video VAE compresses frames, an audio VAE compresses the spectrogram. In LTX-2, one second of sound becomes about twenty-five latent vectors of 128 numbers. Now sound is a short sequence of tokens, the same kind of object the video model already understands.

中文讲解：原理部分，改编自卡内基梅隆大学生成式AI课程第24讲（音频）。原始声音是压力波，每秒采样数万个数字（录音室为48000）。我们先对重叠的短窗口做傅里叶变换，把声音变成一张“音高随时间变化”的图，即频谱图；梅尔频谱再按人耳感知调整音高轴。然后像视频VAE压缩画面一样，音频VAE压缩频谱图：在LTX-2中，一秒声音变成约25个128维潜向量。于是声音也变成了transformer能处理的token序列。

16. See it: 8 seconds of the Veo 3 sailor, as a model sees it

Here is that pipeline run on a real clip: the Veo 3 sailor we just heard. On top, the raw waveform: sixteen thousand numbers per second, so eight seconds is a hundred and twenty-eight thousand numbers. Below, the log-mel spectrogram: a hundred frames per second, eighty bands from low to high pitch. You can read it. The horizontal stripes are the harmonics of his voice on vowels. The vertical bursts are consonants and splashes. The noisy texture between words is the wind and the sea. For art students: this is why audio models borrow image-model ideas. A spectrogram is an image of sound.

中文讲解：把流程用在一个真实片段上：刚才听到的Veo 3老水手。上方是原始波形，每秒16000个数字，8秒就是12.8万个；下方是对数梅尔频谱图，每秒100帧、80个从低到高的频带。它是可以“读”的：水平条纹是元音的谐波，竖直的爆发是辅音和浪花，词间的嘶嘶声是风和海。这也是音频模型借用图像模型思路的原因：频谱图就是声音的图像。

17. Worked example: 10 seconds of picture vs sound

A worked example with real numbers, in the same style as last week's pixels-to-tokens slide. Ten seconds of 720p video at twenty-four frames per second is two hundred and forty frames. The video VAE compresses four times in time, so about sixty-one latent frames. Each latent frame becomes three thousand six hundred tokens, which we computed last week, so about two hundred and twenty thousand video tokens. The same ten seconds of sound, after LTX-2's audio VAE, is two hundred and fifty tokens. Roughly eight hundred and eighty times fewer. That is why LTX-2 gives picture a fourteen-billion-parameter stream and sound a five-billion one. Picture is dense in space; sound is dense in time.

中文讲解：用真实数字算一遍（与上周“像素到token”同一风格）：10秒720p、24帧/秒是240帧；视频VAE时间上压缩4倍，约61个潜帧；每个潜帧3600个token（上周算过），约22万个视频token。同样10秒的声音经过LTX-2音频VAE只有250个token，少约880倍。所以LTX-2给画面140亿参数的流、给声音50亿参数的流。画面在空间上密集，声音在时间上密集。

18. Four ways to wire sound and picture

There are four ways to wire sound and picture together, and every model you'll meet is one of them. One: picture first, then a second model listens to the video and adds sound. That's MMAudio, or Tencent's HunyuanVideo-Foley. Good for Foley, but the picture can never react to the sound. Two: sound first, then animate a face or body to it. That's the talking-head family, story two. Three, the most popular open design this year: two towers, one for video and one for audio, denoising at the same time and passing messages to each other in every block. Ovi, LTX-2 and Kandinsky 6 work this way. Four: throw all the tokens into one sequence and let a single transformer attend to everything. MiniMax H3 and daVinci-MagiHuman do this. It's simpler, but you can't give the two modalities different sizes.

中文讲解：连接声音与画面有四种方式：一、先画面后声音，第二个模型“听”视频再配音，如MMAudio、腾讯混元Foley，拟音好但画面无法响应声音；二、先声音后画面，即说话人模型家族（第二个故事）；三、今年最流行的开源设计：视频塔和音频塔同时去噪，每层互相交换信息，如Ovi、LTX-2、Kandinsky 6；四、所有token放进一个序列由一个transformer处理，如MiniMax H3和daVinci-MagiHuman，更简单但两种模态无法分配不同大小。

19. Inside LTX-2: two streams, one clock

Let's open up LTX-2, because its weights and paper are public. Two streams. The video stream is big, about fourteen billion parameters, and every video token knows where it is in space and time: x, y and t. The audio stream is smaller, five billion, and its tokens only know time. In every block, the two streams look at each other through cross-attention in both directions: video asks audio 'what is happening now?', audio asks video 'what should I sound like?'. And here's the clever bit: when they look at each other, only the time positions are used. A video token at two seconds attends most naturally to the audio at two seconds. That shared clock is how a door slam lands on the right frame. The paper reports up to twenty seconds with stereo sound, and about one point two seconds per diffusion step on an H100, which is eighteen times faster per step than Wan 2.2.

中文讲解：我们打开LTX-2看看（它的权重和论文都是公开的）。两条流：视频流约140亿参数，每个视频token知道自己的空间和时间位置（x、y、t）；音频流约50亿参数，token只知道时间。每一层中，两条流通过双向交叉注意力互相“看”：视频问音频“现在发生了什么”，音频问视频“我应该听起来像什么”。巧妙之处在于，互看时只使用时间位置：2秒处的视频token自然对应2秒处的音频。这个共享时钟让关门声落在正确的帧上。论文称可生成最长20秒的立体声视频，在H100上每步约1.2秒，比Wan 2.2快18倍。

20. From the technical report: LTX-2 at a glance

This is the real overview figure from the LTX-2 report, so you can see that my simplified diagram was honest. Read it left to right. Audio is turned into a mel spectrogram and compressed by an audio VAE; video is compressed by a video VAE. Text goes through a large language model, Gemma 3, and then through two small 'connectors', one for each stream. In the middle, the audio stream and video stream are stacked N times with the cross-attention arrows between them. At the end there are two losses, one for audio and one for video, trained together.

中文讲解：这是LTX-2报告中的原始总览图，可以验证我刚才的简化示意。从左到右：音频转成梅尔频谱后由音频VAE压缩，视频由视频VAE压缩；文本经过大语言模型Gemma 3，再经过两个小“连接器”分别送入两条流；中间音频流与视频流堆叠N层，之间是交叉注意力箭头；最后有音频和视频两个损失，一起训练。

21. Keeping time: why positions must be aligned

One more detail for the technical students, because it's the key to sync. Video tokens and audio tokens arrive at different rates. If you number them naively, token 30 of the video and token 30 of the audio are at different moments, so attention looks at the wrong sound. These two figures from MMAudio, at CVPR 2025, show the attention affinity between visual and audio tokens. On the left, default positions: the bright band drifts. On the right, positions scaled to real time: a clean diagonal, each frame paired with its own moment. LTX-2's 'temporal RoPE only' and Ovi's 'scaled RoPE' are the same idea.

中文讲解：再讲一个同步的关键细节。视频token和音频token的速率不同，若简单编号，第30个视频token与第30个音频token并不在同一时刻，注意力会看错声音。这两张图来自CVPR 2025的MMAudio，显示视觉与音频token的亲密度：左边默认位置，亮带逐渐偏移；右边按真实时间缩放位置，得到干净的对角线，每帧对应自己那一刻的声音。LTX-2的“只用时间RoPE”和Ovi的“缩放RoPE”是同一思路。

22. LTX-2: give it a sound, get a performance

Here's LTX-2 run in the other direction: the input is an audio file, a scream, and the model generates a face and body performing it. [Play.] Because the model has both streams, the same weights can go text to audio and video, image to audio and video, or audio to video, depending on which stream you hold fixed. That flexibility is a big reason open two-stream models are popular with artists: you can bring your own voice track.

中文讲解：这是LTX-2的反向用法：输入一段尖叫音频，模型生成与之匹配的面部和身体表演。由于有两条流，同一组权重可以做文本到音视频、图像到音视频或音频到视频，取决于固定哪一条流。这种灵活性是双流开源模型受艺术家欢迎的重要原因：你可以带自己的声音进来。

23. Ovi: two voices, one take

Ovi, from Character.AI with Yale. [Play.] Two people, two voices, turn-taking, a car interior. The dialogue was written in the prompt; the model decided the timing, the voices and the faces. Ovi's design is the symmetric version of two towers: both towers have the same shape, the video one starts from Wan 2.2, and the audio one was trained from scratch on hundreds of thousands of hours of audio. It's open under Apache 2.0, so it's a good one for your assignments.

中文讲解：Character.AI与耶鲁大学的Ovi：两个人、两种声音、轮流说话、车内环境。对白写在提示词里，时机、声音和面孔由模型决定。Ovi采用对称双塔：两个塔结构相同，视频塔来自Wan 2.2，音频塔在数十万小时音频上从零训练。Apache 2.0开源，适合用于作业。

24. Ovi's twin towers

Here's Ovi's own figure. Two towers of exactly the same shape: video on the left, initialised from Wan 2.2, audio on the right, trained from scratch on a huge amount of audio. Because the towers match block for block, each fusion block can do self-attention, then cross-attention to the text, then cross-attention to the other tower. Notice the snowflake and flame icons: frozen parts versus trained parts. A practical lesson for anyone building models: reuse what already works and train only what's new.

中文讲解：Ovi的原图：两座完全同形的塔，左边视频（从Wan 2.2初始化），右边音频（在海量音频上从零训练）。逐块对应，所以每个融合块都能先自注意、再与文本交叉注意、再与另一座塔交叉注意。注意雪花和火焰图标：冻结部分与训练部分。实用经验：复用已有能力，只训练新的部分。

25. Recipe: how to teach a video model to hear

How do you actually build one? The Kandinsky 6.0 report, published this month, spells out a recipe that most teams now follow. Step one: don't start from scratch; reuse a strong video model. Step two: train the new audio tower alone first, on text-to-audio, so it learns what the world sounds like. Kandinsky used about forty million audio tracks. Step three: join the towers and train them together on video clips that have meaningful sound, about seven million segments after filtering. Step four: polish, with fine-tuning, reinforcement learning, and distillation down to ten steps so it's fast. Baidu's NAVA follows a similar path from Wan 2.2.

中文讲解：怎么构建？本月发布的Kandinsky 6.0报告给出了多数团队采用的配方：一、复用强大的视频模型；二、先单独训练音频塔做文本到音频，学会世界的声音（约4000万条音轨）；三、两塔合并，在有意义声音的视频片段上联合训练（过滤后约700万段）；四、微调、强化学习，并蒸馏到10步以加速。百度NAVA从Wan 2.2出发走了类似路线。

26. Data: finding clips that are really in sync

The unglamorous truth is that data is the hardest part. Most online video is bad training data for sound: music pasted over vlogs, dubbed films, voice-overs where the speaker isn't on screen. Teams use a sync score, a model that checks whether lips match speech, to throw those clips away. Kandinsky keeps a clip with a person on screen only if its sync score is high. Second, captions now describe the sound, not just the image, with spoken words in quotation marks, which is why your prompts should do the same. Third, multi-speaker scenes need to know who is talking: NAVA attaches a voice fingerprint to each line of speech so two characters keep two different voices. For scale, NAVA trained on about fifteen million clips using a hundred and sixty H100 GPUs for about four weeks.

中文讲解：数据是最难的部分。网上大多数视频不适合训练声音：后贴音乐的vlog、配音电影、画外音。团队用“同步分数”模型剔除口型不匹配的片段。其次，字幕现在也描述声音，对白放在引号里，所以你的提示词也应如此。第三，多人场景要知道谁在说话：NAVA为每句台词绑定说话人的声纹，让两个角色保持两种声音。规模上，NAVA用约1500万片段、160张H100训练了约四周。

27. Anatomy of an audio-video prompt

Practically, how do you prompt for sound? Think in five layers, like a sound mixer does. Picture, as before. Dialogue: who speaks, the exact words in quotation marks, and how they say it. Effects: sounds tied to visible actions, the wipers, the turn signal. Ambience: the always-on bed of the place, rain on the roof. And music: decide whether it's inside the scene, a radio in the car, or a score over the scene, and keep it under the voice. Two practical tips from people who use Seedance a lot: keep each spoken line short, and don't ask

for big head turns while someone is talking, because motion and lip-sync compete.

中文讲解：实操上如何为声音写提示？像混音师一样分五层：画面；对白（谁说、引号中的原话、怎么说）；音效（与可见动作绑定的声音）；环境声（场所持续的底噪）；音乐（是场景内的收音机还是场景外的配乐，压在人声下面）。

Seedance用户的两条经验：台词要短，说话时不要要求大幅转头，因为动作和口型会互相干扰。

28. Copy these: three prompts that work

Here are three prompts you can copy tonight. A dialogue scene: notice the exact words in quotation marks, how they're said, one effect tied to an action, the ambience, and 'No music', because otherwise many models add a score. A sound-led shot with no speech, where every sound is tied to a visible action. And music inside the scene, where we say the music is played on screen, and where the camera should be on the last note. Same five layers as the table.

中文讲解：三个今晚就能复制使用的提示：对白场景——注意引号里的原话、说话方式、一个与动作绑定的音效、环境声，以及“不要音乐”（否则很多模型会加配乐）；以声音为主、无对白的镜头——每个声音都绑定可见动作；场景内音乐——说明音乐在画面中演奏、最后一个音符时镜头停在哪里。与上表的五层一致。

29. Be the Foley artist: what should this sound like?

Second game. You are the Foley artist now. [Play muted.] Write down every sound you would put in this shot, in order. Wheels on asphalt, leaves, the board hitting a crack, wind? Thirty seconds. [Play with sound.] Now compare. What did the model include that you forgot? What did you include that it didn't? In my experience, models are good at the continuous bed, like wheels and ambience, and less reliable on the exact, single events. That's where a human sound designer still adds the most.

中文讲解：第二个游戏：你是拟音师。先静音播放，写下你会为这个镜头加的所有声音：轮子与路面、落叶、滑板撞到裂缝、风声？30秒后打开声音对比。模型加了什么你忘了？你加了什么模型没有？经验上，模型擅长连续的底层声音，对单个精确事件不太可靠，这正是人类声音设计师最能发挥的地方。

30. Listen twice: which Foley is better?

Third game, a blind A/B test. Same burning grassland, same picture, sound added afterwards by two different video-to-audio models. [Play A, then B.] Hands up for A. Hands up for B. Why? The reveal: A is Tencent's HunyuanVideo-Foley from 2025, B is MMAudio from the year before. Tencent's paper reports better synchronization and semantic match for A on their benchmark, but notice the question you just answered is not 'which is more accurate' but 'which do I believe'. That's synchresis again, and it's what next week's evaluation lecture is about.

中文讲解：第三个游戏：盲听A/B。同一片燃烧的草原、同一画面，由两个不同的视频转音频模型事后配音。先播A再播B，举手投票并说明理由。揭晓：A是腾讯2025年的HunyuanVideo-Foley，B是前一年的MMAudio。腾讯论文报告A在同步和语义匹配上更好，但你刚才回答的不是“哪个更准确”，而是“我相信哪个”——这又是合成同步，也是下周评估课的主题。

31. Measure it: do the sounds line up with the motion?

Can we measure sync without a neural network? Yes, roughly, and I did it on the wok clip for tonight. The green line is how much each video frame changes: motion. The orange line is how much new energy appears in the spectrogram: sound onsets. Many peaks coincide: the toss and the clatter happen together. This is the simplest possible version of what models like Synchformer learn to do.

中文讲解：不用神经网络能否测量同步？可以粗略测量，我用炒锅片段做了：绿线是每帧画面变化量（运动），橙线是频谱中新出现的能量（声音起始）。很多峰值重合：翻锅和碰撞声同时发生。这是Synchformer这类模型所学内容的最简版本。

32. Slide the sound: a hand-made SyncNet

Now slide the sound against the picture in ten-millisecond steps and score the match at each shift. The best match is about forty milliseconds from zero, one frame at twenty-four frames per second, well inside the range the ITU found viewers can't detect. That is exactly what SyncNet does, except SyncNet learns the

features instead of using my hand-made ones. Keep this picture in mind for story two.

中文讲解：把声音以10毫秒为步长相对画面滑动，并在每个位移处计算匹配度。最佳匹配约在40毫秒，相当于24帧下的一帧，远在观众察觉不到的范围内。SyncNet做的正是这件事，只是它学习特征，而不是用我手工设计的特征。第二个故事会再用到这张图。

33. Diegetic or not? Music that lives in the scene

A term from film studies that's very useful when prompting: diegetic sound comes from inside the story world; the characters could hear it. Non-diegetic sound, like a film score, is only for the audience. [Play both.] On the left, Ovi's street saxophonist: diegetic. On the right, UniVerse-1, a 2025 model that stitched together a pretrained video model and a pretrained music model, plays a guitarist. Joint models are very good at diegetic music because they can see the instrument. They are much less controllable for score, so many filmmakers still add the score themselves.

中文讲解：电影研究中一个对写提示很有用的术语：剧情内声音（diegetic）来自故事世界内部，角色也能听到；剧情外声音（如配乐）只给观众听。左边是Ovi生成的街头萨克斯手（剧情内），右边是2025年的UniVerse-1，它把预训练视频模型和音乐模型“拼接”在一起，生成弹吉他的人。联合模型很擅长剧情内音乐，因为它能“看到”乐器；对配乐的可控性要差得多，所以很多导演仍自己加配乐。

34. Where it still breaks

Where does it still break? Six things to watch for in your own work. Two people talking: the wrong mouth moves, or both do. ByteDance itself lists multi-person lip-sync as an open problem. Motion versus mouth: a head turn or a laugh in the middle of a sentence and the lips lose track. Audio artefacts: metallic voices, clipped words, music swelling over dialogue. Language: quality varies a lot across languages and dialects. Exact events: the crash lands a few frames early, and thanks to synchresis, you notice. And voice continuity: the same character can sound different in the next shot, unless you give the model a voice reference, which is where story two begins.

中文讲解：仍然会出错的地方：两人对话时错误的嘴在动或两张嘴都在动（字节自己也承认这是未解决问题）；说话时转头或大笑导致口型跟不上；音频瑕疵：金属音、吞字、音乐盖过对白；不同语言和方言的口型质量差异很大；精确事件提前几帧（因为合成同步你会注意到）；同一角色在下一个镜头声音变了，除非给声音参考——这正是第二个故事的起点。

35. Can you run it yourself? Speed, again

Can you run it yourself? The numbers are moving fast. LTX-2 is about eighteen times faster per step than Wan 2.2 at similar size, because the audio stream is small and the video compression is aggressive. And in March, a team from JD, Fudan, Peking University and HKU published OmniForcing, which distills LTX-2 into a streaming model: a five-second clip with sound in under six seconds instead of three minutes, about twenty-five frames per second on one GPU. The tricks are exactly what Yaofu and I showed you: fewer steps by distillation, chunks with a KV cache, and quantization. Keep that in mind for story three.

中文讲解：自己能跑吗？速度提升很快。LTX-2每步比同规模的Wan 2.2快约18倍，因为音频流小、视频压缩率高。今年3月，京东、复旦、北大和港大团队发表的OmniForcing把LTX-2蒸馏成流式模型：5秒带声音的视频不到6秒生成（原来要3分钟），单卡约25帧/秒。用的正是耀夫和我讲过的技巧：蒸馏减少步数、分块加KV缓存、量化。第三个故事会再用到。

36. For sound designers and composers: what changes?

For the sound designers and composers in the room: what changes? The model now gives you a first-draft mix that's already in sync, with room tone and footsteps for free, diegetic music that matches the instrument on screen, and temporary dialogue for pitching. What stays human? Sound design was never about adding every sound; it's about choosing what not to hear, where the silence goes, whose point of view we're in. The exact event, the hit on the cut. The score, the emotion the characters can't hear. And the decision about whose voice is used, which is story four.

中文讲解：对在座的声音设计师和作曲者：模型现在能给出已同步的初版混音、免费的房间声和脚步声、与画面乐器匹配的剧情内音乐，以及用于提案的临时对白。仍属于人的：声音设计从来不是加入所有声音，而是选择不让人听到什么——沉默放在哪里、站在谁的视角；精确的事件和剪辑点；配乐（角色听不到的情绪）；以及使用谁的声音的决定（第四个故

事)。

37. If the model writes the sound with the picture, who is the sound designer now?

For the room, two minutes with your neighbour. If the model writes the sound together with the picture, who is the sound designer? The person who wrote ‘rain on the roof’ in the prompt? The model? Or the thousands of Foley artists and recordists whose work trained it? There’s no right answer yet, but you’ll need one for your assignment credits.

中文讲解：和邻座讨论两分钟：如果模型和画面一起生成声音，谁是声音设计师？写下“雨打车顶”的提示者？模型？还是训练数据中成千上万的拟音师和录音师？目前还没有标准答案，但你的作业署名需要一个答案。

38. Directing a performance, not a prompt

Story two. Once a model can make a character speak, the question changes. It’s no longer ‘do the lips match the audio?’ It becomes the question every director asks: is this a good performance, and how do I get a better one? This story walks through how talking-head models grew into acting models, how we measure sync, and how you direct them.

中文讲解：第二个故事。一旦模型能让角色说话，问题就变了：不再是“嘴型对不对”，而是每个导演都会问的问题：这是好的表演吗？如何得到更好的表演？我们会看说话人模型如何发展成表演模型、如何衡量同步，以及如何导演它们。

39. What is a performance made of?

What is a performance made of? Let’s break it into six parts. Voice: pitch, pace, breath, accent. Timing: pauses, interruptions, the beat before the punchline. Mouth: lips, jaw and teeth matching each sound. Face and eyes: brows, blinks, gaze. Body: head motion, shoulders, hands. And listening, what you do while the other person is talking, which actors will tell you is half of acting. The first models only did the mouth. Today’s models attempt all six, and listening is the newest, which is story three.

中文讲解：表演由什么组成？分成六部分：声音（音高、语速、呼吸、口音）；时机（停顿、打断、笑点前的一拍）；嘴（唇、下颌、牙齿与每个音匹配）；面部与眼睛（眉毛、眨眼、视线）；身体（头部、肩膀、手）；以及倾听——对方说话时你在做什么，演员会说这是表演的一半。早期模型只做嘴，如今的模型尝试全部六项，倾听是最新的（第三个故事）。

40. From lip-sync to acting

The lineage, quickly. In 2020, Wav2Lip repaints only the mouth of an existing video. In April 2024, Microsoft’s VASA-1 animates a whole face and head from one photo at forty frames per second. In February 2025, ByteDance’s OmniHuman-1 does the full body, gestures and singing, trained on over eighteen thousand hours of human video. In mid-2025, MultiTalk and InfiniteTalk from the MeiGen-AI team handle two people and unlimited length. Then in August 2025, OmniHuman-1.5 does something that sounds like a director: a language model first reads the audio and writes a plan for the acting, and the video model follows the plan. And in 2026, with native audio-video, the voice itself is generated together with the face.

中文讲解：简要谱系：2020年Wav2Lip只重绘嘴部；2024年4月微软VASA-1用一张照片驱动整张脸和头部，40帧/秒；2025年2月字节OmniHuman-1生成全身、手势和唱歌，训练数据超过1.87万小时；2025年中MultiTalk和InfiniteTalk处理两人对话和无限时长；2025年8月OmniHuman-1.5像导演一样：先由语言模型读懂音频、写出表演计划，再由视频模型执行；2026年原生音视频模型连声音也和脸一起生成。

41. Sounds you can see: phonemes → visemes

Before the machine learning, some phonetics, because it helps you judge results with your own eyes. Sounds are phonemes; the mouth shapes you can see are called visemes, and many sounds share one viseme. The ones that matter most are the bilabials, p, b and m, where the lips must fully close. If a character says ‘Bob’ and the lips never meet, everyone notices. F and v, where the teeth touch the lower lip, are hard for models. Open vowels are easy. And tongue sounds like t, d, k are mostly hidden, so models get them for free. When you review a generated performance, watch the p, b and m sounds. That’s the fastest quality check there is.

中文讲解：先讲一点语音学，帮助你用眼睛判断结果。声音单位叫音素，能看到的嘴型叫视位，多个音素可共享一个视位。最关键的是双唇音p、b、m，嘴唇必须完全闭合——角色说“Bob”嘴唇却没合上，人人都会发现。f、v（上齿碰下唇）对模型较难；开口元音容易；t、d、k等舌音大多藏在嘴里，模型“白得”。检查生成表演时，盯住p、b、m，这是最快的质量检查。

42. Under the hood: how a voice drives a face

How does a voice actually drive a face? The audio goes into a pretrained speech model, usually wav2vec 2.0 from Meta or OpenAI's Whisper. These models were trained to understand speech, so their features already encode which phoneme is being spoken, and a bit of emotion. We resample those features so there is one vector per video frame. Then inside the video transformer, every frame's tokens use cross-attention to ask 'what sound is happening right now?' A reference photo keeps the identity. VASA-1 and OmniHuman use wav2vec; LatentSync and MuseTalk use Whisper. The nice idea here is reuse: the video model doesn't have to learn phonetics from scratch, it borrows it from a speech model.

中文讲解：声音如何驱动面孔？音频先输入预训练语音模型（通常是Meta的wav2vec 2.0或OpenAI的Whisper），其特征已编码了当前的音素和一些情绪。把特征重采样到每个视频帧一个向量，在视频transformer中，每帧的token通过交叉注意力询问“此刻是什么声音”，参考照片保持身份。VASA-1和OmniHuman用wav2vec，LatentSync和MuseTalk用Whisper。核心是复用：视频模型不必从零学习语音学。

43. Under the hood: how we measure lip-sync (SyncNet)

How do papers measure lip-sync? Almost all of them use the same tool: SyncNet, from Oxford in 2016. Take five frames of the mouth, a fifth of a second, and the matching fifth of a second of audio. SyncNet turns both into vectors, trained so that matching pairs are close and mismatched pairs are far apart. Now slide the audio backwards and forwards, up to fifteen frames each way, and find the shift where the distance is smallest. That smallest distance is Sync-D, lower is better. Sync-C is the confidence: how much that best shift stands out from all the wrong ones, higher is better. If the best shift isn't zero, the clip is out of sync by that many frames.

中文讲解：论文如何衡量口型同步？几乎都用牛津2016年的SyncNet：取5帧嘴部（0.2秒）和对应0.2秒音频，SyncNet把两者变成向量，同步时距离近、不同步时距离远。把音频前后滑动最多15帧，找到距离最小的位置：最小距离即Sync-D（越低越好）；Sync-C是置信度，即最佳位置比其他位置突出多少（越高越好）。若最佳偏移不是0，说明片段错位了那么多帧。

44. Reading the scoreboard (and its trap)

Here's how to read the scoreboard. On the lip-reading dataset LRS2, real video scores about 7.8 Sync-C and 6.7 Sync-D. Wav2Lip, from 2020, scores 7.8 and 6.4, a better Sync-D than reality. Did it beat reality? No. Wav2Lip was trained against a SyncNet-style judge, so it learned to please that judge. That's a general lesson for next week's lecture on evaluation: when a metric becomes a training target, it stops measuring what you care about. Below, on the newer Verse-Bench, the open joint models: NAVA from Baidu leads, then Ovi, LTX 2.3 and daVinci-MagiHuman, all close together. Only compare numbers within the same benchmark.

中文讲解：如何读排行榜：在唇读数据集LRS2上，真实视频Sync-C约7.8、Sync-D约6.7；2020年的Wav2Lip为7.8和6.4，Sync-D比真实视频还好。它超越现实了吗？没有。Wav2Lip训练时以类似SyncNet的判别器为“裁判”，学会了讨好这个裁判。这是下周评估课的通用教训：指标一旦成为训练目标，就不再衡量你关心的东西。下方是较新的Verse-Bench上的开源联合模型：百度NAVA领先，其次是Ovi、LTX 2.3和daVinci-MagiHuman，差距很小。只在同一基准内比较数字。

45. Game: which one is in sync?

A game to feel the numbers. Three versions of the same sailor clip. I re-timed two of them. [Play A, B, C.] Vote: which one is in sync? The answer: B is the original. In A, the sound is 250 milliseconds late, six frames, and most of you probably saw it. In C, the sound is 120 milliseconds early, about three frames. Early sound is much easier to notice than late sound, which matches the ITU numbers: about 45 milliseconds early versus 125 late. Why the asymmetry? In real life, light arrives before sound, so our brains are used to late sound and

suspicious of early sound.

中文讲解：一个体验数字的游戏：同一老水手片段的三个版本，其中两个我重新调整了时间。投票：哪个是同步的？答案：B是原版；A的声音晚250毫秒（6帧），大多数人应该能看出；C的声音早120毫秒（约3帧）。声音提前比滞后更容易被察觉，与国际电信联盟的数据一致（提前约45毫秒、滞后约125毫秒）。为什么不对称？现实中光比声音先到，大脑习惯声音晚到，对提前的声音更敏感。

46. VASA-1: a painting that talks

Microsoft's VASA-1 from 2024. [Play.] The Mona Lisa, rapping. The trick is a learned face space that separates who you are from how you move: identity and appearance on one side, head pose and facial dynamics on the other. A surprisingly small diffusion transformer, about twenty-nine million parameters, generates only the motion, which is why it can run at forty frames per second on one RTX 4090 with about 170 milliseconds of delay. Microsoft never released the model or an API, explicitly because of the risk of impersonation. Remember that when we get to story four.

中文讲解：微软2024年的VASA-1：会说唱的蒙娜丽莎。关键是一个学到的面部空间，把“你是谁”（身份与外观）和“你怎么动”（头部姿态与面部动态）分开。一个仅约2900万参数的扩散transformer只生成动作，因此在一张RTX 4090上能达到40帧/秒、约170毫秒延迟。微软明确出于冒充风险没有发布模型或API——第四个故事会再提到。

47. How VASA-1 works: generate motion, not pixels

VASA-1's own figure, and the idea is worth stealing. On the left, training: the model learns to generate a motion latent, a compact code for head pose and face dynamics, from audio features. On the right, generation: your photo is encoded once into appearance and identity, then only the motion is generated, and a decoder turns appearance plus motion into frames. Because the motion code is tiny compared with pixels, it can run at forty frames per second. Generate the small thing, render the big thing.

中文讲解：VASA-1的原图，思路值得借鉴。左边是训练：模型根据音频特征生成“运动潜变量”（头部姿态与面部动态的紧凑编码）；右边是生成：照片只编码一次为外观和身份，然后只生成运动，再由解码器合成画面。运动编码远小于像素，所以能达到40帧/秒。生成小的，渲染大的。

48. Same face, same words, four emotions

Here is why VASA-1 matters for directing. [Play.] Same photo, same audio, four different emotional settings. Because motion is separated from identity, emotion, gaze direction and head distance become separate knobs. That's the beginning of directing a performance rather than generating one. Ask yourself as you watch: which one would you cast?

中文讲解：VASA-1对导演的意义：同一张照片、同一段音频、四种情绪设置。因为动作和身份是分开的，情绪、视线方向和头部距离成为独立的“旋钮”。这就是从“生成表演”到“导演表演”的开始。看的时候问自己：你会选哪一个？

49. OmniHuman: from face to whole body

ByteDance's OmniHuman. [Play.] Now the whole body performs: hands on the keys, breathing, posture. The training trick is to mix conditions: some clips come with only text, some with audio, some with pose, so the model can use eighteen thousand hours of messy human video instead of a small clean set. Version 1.5, last August, adds something that looks like a director's note: a multimodal language model first listens to the audio and writes a structured plan, what emotion, what gesture, when, and the video model follows it.

中文讲解：字节的OmniHuman：现在是全身表演——手在琴键上、呼吸、姿态。训练技巧是混合条件：有的片段只有文本，有的有音频，有的有姿态，从而能使用1.87万小时杂乱的人类视频。去年8月的1.5版加入了类似“导演笔记”的东西：多模态大模型先听音频、写出结构化计划（什么情绪、什么手势、何时），视频模型再照做。

50. OmniHuman-1.5: acting, not just talking

OmniHuman-1.5 in a dramatic scene. [Play.] Don't watch the lips this time; they're fine. Watch the eyes, the pauses, where he looks away. That's the part the language-model plan is supposed to add: intention. Is it convincing? Where does it feel generic?

中文讲解：OmniHuman-1.5的一段戏剧场景。这次不要看嘴，嘴没问题；看眼睛、停顿、他移开视线的时刻。这正是语

言模型计划要增加的部分：意图。它有说服力吗？哪里显得套路化？

51. OmniHuman-1.5: a planner thinks, a renderer acts

And here's how OmniHuman-1.5 adds intention. The authors borrow the psychology idea of two systems. System two, slow thinking: multimodal language-model agents listen to the audio, look at the image, read the text, and write a plan: emotion, action, shot, for each moment. System one, fast acting: a diffusion transformer with text, audio and video branches renders the plan. It's literally a director and an actor.

中文讲解：OmniHuman-1.5如何加入“意图”：作者借用心理学的“双系统”。系统二（慢思考）：多模态大模型智能体听音频、看图像、读文本，为每个时刻写出情绪、动作和镜头计划；系统一（快行动）：带文本、音频、视频分支的扩散transformer按计划渲染。字面意义上的导演与演员。

52. Singing: voice drives breath, body and instrument

Singing is harder than speech: long vowels, breath, the body moving with the rhythm, and often an instrument. [Play both.] Left, Alibaba's Wan2.2-S2V, built for film: a singer on a ship in a storm, with the camera and body moving with the song. Right, Tencent's HunyuanVideo-Avatar: strumming hands and breathing from one photo and a song track. Watch the hands: are they playing the right rhythm?

中文讲解：唱歌比说话难：长元音、呼吸、身体随节奏摆动，还常有乐器。左边是面向影视的阿里Wan2.2-S2V：风暴中船上的歌手，镜头和身体随歌声运动；右边是腾讯HunyuanVideo-Avatar：一张照片加一首歌生成弹奏的手和呼吸。看看手：节奏弹对了吗？

53. Two people: who is speaking?

Two people is much harder than one, because the model must decide which mouth each voice belongs to. [Play both.] MultiTalk, on the left, takes one audio track per person and learns to bind each track to the right face; Tencent's HunyuanVideo-Avatar, on the right, does duets and adds an emotion module. Listen to the turn-taking. Remember that ByteDance still calls multi-person lip-sync an open problem.

中文讲解：两个人比一个人难得多，因为模型要判断每个声音属于哪张嘴。左边MultiTalk每人一条音轨并学会绑定到正确的脸；右边腾讯HunyuanVideo-Avatar能做对唱并加入情绪模块。注意听轮流说话。字节仍把多人口型同步列为未解决问题。

54. Dubbing: one performance, many languages

One of the most immediate uses is dubbing. [Play.] InfiniteTalk takes an existing video and a new voice track and regenerates not just the lips but the head, body and expression, for any length. Hollywood got here first: the 2022 thriller Fall used a company called Flawless to change more than thirty swear words so the film could get a PG-13 rating, without reshooting. They call it 'vubbing'. Ask yourself: when an actor's face now speaks Japanese, whose performance is it?

中文讲解：最直接的应用之一是配音。InfiniteTalk输入现有视频和新的配音，不仅重新生成嘴型，还生成头部、身体和表情，时长不限。好莱坞更早：2022年的惊悚片《坠落》用Flawless公司的技术修改了30多处脏话以获得PG-13评级，无需重拍，称为“vubbing”。问问自己：当演员的脸开始说日语，这是谁的表演？

55. A performance built from references

Here is a clip from Vidu's homepage. [Play.] A fortune teller, candlelight, a voice that suits her. Vidu doesn't say which model version made this one, but its new Q4 model, released two weeks ago, lets you pin the character, the wardrobe, the set and the voice, each with its own reference. That's the direction for all of these tools: from typing a wish to assembling a cast.

中文讲解：这是Vidu主页上的一段：烛光中的算命师，声音与形象相配。Vidu没有说明这段由哪个版本生成，但两周前发布的Q4允许分别用参考固定角色、服装、场景和声音。所有这些工具的方向都是：从输入愿望，到组建演员阵容。

56. Directing with references

So how do you direct? Mostly with references. Identity: Vidu Q4, released two weeks ago, takes up to fifteen

reference images to lock a character, costume and set. Voice: Vidu Q4 and Seedance accept reference audio, so a character sounds the same in shot two as in shot one. The most powerful option for artists: bring your own track. Record the line yourself, with exactly the timing you want, and use an audio-to-video model to act it. And for scenes with several people, tell the model which line belongs to whom. A rule of thumb: the more you specify by reference, the less you leave to chance, and the more consent you need from whoever's face or voice that is.

中文讲解：怎么导演？主要靠参考。身份：两周前发布的Vidu Q4最多接受15张参考图锁定角色、服装和场景。声音：Vidu Q4和Seedance接受参考音频，让角色在不同镜头声音一致。对艺术家最有力的方式：自己录台词，用你想要的节奏，再用音频到视频模型来“演”。多人场景要告诉模型哪句台词属于谁。经验法则：参考给得越多，偶然性越小，需要的授权也越多。

57. Directing the delivery: what to ask for

Here's a practical table for directing delivery. An emotion label alone, like 'angry', gives you the average angry face from the training data. Directors and actors talk instead about playable actions: what the character does. 'Angry but held in' means clipped speech, even pitch, a tight jaw, a fixed stare; you can write 'through clenched teeth, without raising his voice'. 'Nervous' means rising sentence endings, filler words, gaze darting away. Put the behaviour in the prompt, not just the label. This works for actors and, it turns out, for models.

中文讲解：导演台词的实用表：只写“愤怒”只能得到训练数据里平均的愤怒脸。导演和演员讲的是“可表演的动作”——角色在做什么。“压抑的愤怒”是语句短促、音高平稳、下颌紧绷、目光固定，可以写“咬着牙、不提高音量”；“紧张”是句尾上扬、口头语多、目光闪躲。把行为写进提示词，而不只是标签。这对演员有效，对模型也有效。

58. One line, three deliveries: "I never said you could leave."

Exercise, five minutes, in pairs. Here's one line: 'I never said you could leave.' Write three prompts that make it a threat, a joke, and a plea. For each, describe the voice, the face, and one gesture, using playable actions, not labels. If you have Seedance, Kling, Vidu or LTX-2 access on your laptop, try one. I'll ask two pairs to read their prompts aloud.

中文讲解：练习，两人一组，五分钟。台词：“我从没说过你可以走。”写三个提示，分别把它变成威胁、玩笑和恳求。每个都描述声音、表情和一个手势，用可表演的动作而非标签。如果你能用Seedance、可灵、Vidu或LTX-2，就试一个。之后请两组朗读他们的提示。

59. Performers have been digitised before

None of this is entirely new. In 2002, Andy Serkis played Gollum through motion capture, and people argued whether that counted as acting. In 2021, the documentary Roadrunner synthesised three lines in Anthony Bourdain's voice after his death, and the director's casual reply, that they could have an ethics panel about it later, caused an uproar. In 2022, James Earl Jones licensed his archived voice so Darth Vader could keep speaking. Also in 2022, ABBA Voyage opened in London: digital ABBA built by Industrial Light and Magic from about a hundred and sixty cameras of motion capture, more than four million visitors by this year. And last year, The Brutalist used AI to refine the actors' Hungarian vowels, and it became an Oscar controversy. The difference now is scale and cost: what took ILM five weeks takes one photo.

中文讲解：这些并非全新。2002年安迪·瑟金斯通过动作捕捉饰演咕噜，人们争论这算不算表演；2021年纪录片《Roadrunner》在安东尼·波登去世后合成了他的三句话，导演一句“以后再开伦理讨论会”引发轩然大波；2022年詹姆斯·厄尔·琼斯授权他的存档声音让达斯·维达继续说话；同年ABBA Voyage在伦敦开演，工业光魔用约160台摄像机动捕打造数字ABBA，今年观众已超400万；去年《粗野派》用AI修饰演员的匈牙利语音，成为奥斯卡争议。现在的不同在于规模和成本：工业光魔花五周的事，现在只要一张照片。

60. Break: student presentations

Let's pause here for this week's student presentations, and then a ten-minute break. A small challenge for the break: pick any demo from tonight, play it on your phone with the sound off, then on, and notice what changed in how you read the face.

中文讲解：在这里暂停，进行本周学生展示，然后休息十分钟。休息小挑战：在手机上选一个今晚的演示，先静音看再开

声音看，注意你对表情的解读有什么变化。

61. Characters that listen

Welcome back. Story three. Everything so far was offline: you write a prompt, you wait, you get a file. Story three is about characters you can talk to, live. That needs two things the offline models don't have: they must be fast enough, and they must listen, not just speak. This is also where Yaofu's quantization lecture and last week's Self Forcing come back.

中文讲解：欢迎回来。第三个故事。到目前为止都是离线生成：写提示、等待、得到文件。第三个故事讲的是可以实时对话的角色。这需要离线模型没有的两点：足够快，并且会听而不只是会说。耀夫的量化讲座和上周的Self Forcing在这里派上用场。

62. Two numbers that define “real time”

Real time means two different numbers, and people mix them up. Throughput: can you make video at least as fast as it plays? At twenty-five frames per second, that's forty milliseconds per frame. Latency: after you stop talking, how long until the character responds? In human conversation, across ten languages studied, the typical gap between turns is around two hundred milliseconds. Here's the remarkable part: planning a spoken reply takes a human over six hundred milliseconds. So we don't wait for the other person to finish; we predict the end of their sentence and start preparing early. A real-time avatar must do the same. A model can have high throughput and still feel slow, if the first frame takes two seconds.

中文讲解：“实时”包含两个常被混淆的数字。吞吐量：生成速度至少和播放一样快，25帧/秒即每帧40毫秒。延迟：你说完后角色多久回应？对十种语言的研究显示，人类对话的轮换间隔约200毫秒。而人规划一句回答需要600毫秒以上，所以我们会预测对方句子何时结束、提前准备。实时数字人也必须这样。模型可以吞吐量很高，但如果第一帧要两秒，仍会感觉很慢。

63. The latency budget of a talking avatar

Here's where the time goes in a typical talking avatar today, as a pipeline. First the system must decide you've stopped talking, which takes a few hundred milliseconds. Speech recognition, about a hundred. The language model's first words, about four hundred. Text-to-speech, first audio, about a hundred and fifty. Then the video model's first frame, anywhere from under two hundred milliseconds to over a second depending on the model. Plus the network. Run in sequence, that's about one and a half seconds, seven times the human gap. The fixes: overlap everything by streaming, so each stage starts before the previous one finishes; predict the end of the turn; and make the video model work in small chunks. Or, the newest idea, merge the whole pipeline into one model.

中文讲解：当今典型数字人管线的时间分配：先判断你已说完（几百毫秒），语音识别约100毫秒，大模型首个词约400毫秒，语音合成首段约150毫秒，视频模型首帧从不到200毫秒到超过1秒，再加网络。顺序执行约1.5秒，是人类间隔的七倍。解决办法：所有环节流式重叠、预测轮次结束、视频模型分小块生成；或者最新的思路——把整个管线合并为一个模型。

64. Why a normal video diffusion model can't talk back

Why can't a normal video diffusion model just talk back? Three multipliers. A typical sampler denoises thirty to fifty times. Classifier-free guidance, from Week 3, runs each step twice. And worst for conversation: in a bidirectional model every frame attends to every other frame, including the future, so the first frame isn't finished until the last frame is. That's seconds to minutes. Great for a film, useless for a conversation.

中文讲解：普通视频扩散模型为什么不能实时回应？三个乘数：采样通常去噪30到50次；第三周讲的无分类器引导每步要算两次；最致命的是双向模型中每帧都关注所有其他帧（包括未来帧），最后一帧完成前第一帧也不可用。结果是几秒到几分钟，拍电影可以，对话不行。

65. Three tricks you already know

The good news is that you already know the three tricks. One, from last week: generate in small chunks,

looking only backwards, and keep a KV cache of the past so you don't recompute it; Self Forcing trains on the model's own output so errors don't snowball. Two: distillation. Train a fast student to jump to the answer in one to four steps. Character.AI's TalkingMachines uses a method called DMD to go from twenty-four model evaluations to two. Three, from Yaofu's lecture: quantization, eight or four bits, so the same GPU does more work per millisecond. And a systems trick on top: run the transformer and the decoder on different GPUs so they work in parallel.

中文讲解：好消息是，这三个技巧你们都学过：一、上周的分块生成，只看过去，用KV缓存保存历史不重复计算；Self Forcing用模型自己的输出训练以防误差累积。二、蒸馏：训练快速学生模型在1到4步内得到结果，Character.AI的TalkingMachines用DMD把24次模型计算降到2次。三、耀夫讲过的量化：8位或4位，让同一张卡每毫秒做更多。另加一个系统技巧：transformer和解码器放在不同GPU上并行。

66. VASA-1 streaming: a face in a live call (2024)

Here's what streaming looked like already in 2024, with VASA-1. [Play a part.] The interface takes a photo and live audio and the face responds as you speak. Up to forty frames per second at 512 by 512, about 170 milliseconds of delay, on a single gaming GPU. It works because VASA only generates a compact motion code, not pixels from scratch.

中文讲解：2024年VASA-1的流式效果：界面输入一张照片和实时音频，脸随你说话而动。单张游戏显卡上 512×512 、最高40帧/秒、约170毫秒延迟。之所以可行，是因为VASA只生成紧凑的动作编码，而不是从零生成像素。

67. TalkingMachines: a FaceTime call with a character

TalkingMachines from Character.AI, 2025. [Play.] A FaceTime-style call with a cartoon fox. Under the hood: they took Wan 2.1, the open image-to-video model, and turned it into a real-time avatar model. Distillation cuts each chunk from twenty-four model evaluations to two. Video is made in chunks of three latent frames, and the KV cache keeps two things: the very first chunk, which anchors who the character is, and the previous chunk, for continuity. That anchoring trick, keep the first frame forever, appears in almost every streaming avatar paper.

中文讲解：Character.AI 2025年的TalkingMachines：和卡通狐狸视频通话。原理：把开源图生视频模型Wan 2.1改造成实时数字人模型；蒸馏把每块的模型计算从24次降到2次；以3个潜帧为一块生成，KV缓存保留两样：最开始的一块（锚定角色身份）和前一块（保证连续）。“永远保留第一帧”这个锚定技巧几乎出现在所有流式数字人论文中。

68. How streaming distillation works (TalkingMachines)

How do you turn a slow film model into a live one? TalkingMachines' figure shows it. Training clips are cut into chunks of three latent frames, and the reference image is always chunk zero. Each chunk is only allowed to look at itself, the first chunk and the previous chunk, a sparse causal mask. Then a slow bidirectional teacher, twenty-four steps, supervises a fast causal student that uses two. And as in last week's Self Forcing, the training data can include the student's own generated frames, so it learns to recover from its own mistakes.

中文讲解：如何把慢速电影模型变成实时模型？TalkingMachines的图展示了方法：训练片段切成每块3个潜帧，参考图像永远是第0块；每块只能看自己、第一块和前一块（稀疏因果掩码）；一个24步的双向慢老师监督一个2步的因果快学生；和上周的Self Forcing一样，训练数据可包含学生自己生成的帧，学会从自身错误中恢复。

69. Where the milliseconds go: split the work across GPUs

And the systems trick from the earlier slide, in their own figure. Case one: a single GPU denoises a chunk, decodes it with the VAE, then sits idle. Case three: one GPU decodes while another already denoises the next chunk, a pipeline, like an assembly line. Case four adds more GPUs for the diffusion steps themselves. The red line, time to first frame, moves earlier each time. For engineers: real time is as much scheduling as modelling.

中文讲解：前面提到的系统技巧，用他们自己的图说明：情况一，单卡先去噪、再VAE解码、然后空闲；情况三，一张卡解码的同时另一张卡已在为下一块去噪，像流水线；情况四，用更多卡并行做扩散步骤。红线（首帧时间）每次都更早。对工程师而言：实时既是建模问题，也是调度问题。

70. StreamAvatar: the whole upper body, live

StreamAvatar, from Tencent Hunyuan and university partners, goes beyond the head: gestures and the upper body, live. [Play both.] They distilled the open Wan 2.2 model into a three-step streaming student: twenty-five frames per second with about 1.2 seconds of delay, on two H800 GPUs. And like TalkingMachines, it keeps the first frames permanently in memory, which they call a reference sink, so the person doesn't slowly turn into someone else.

中文讲解：腾讯混元与高校合作的StreamAvatar超越了头部：实时生成手势和上半身。他们把开源Wan 2.2蒸馏成3步流式学生模型，在两张H800上达到25帧/秒、约1.2秒延迟。和TalkingMachines一样，它把最初几帧永久保留在记忆中（称为“参考锚点”），避免人物逐渐变成另一个人。

71. Avatar Forcing: an avatar that listens (CVPR 2026)

The newest piece is listening. Avatar Forcing, from KAIST and NTU Singapore, at CVPR this year. It takes your audio and your head motion as inputs, so while you're talking, the avatar nods, smiles, or laughs at your joke. Technically it's diffusion forcing: it generates ten-frame blocks with a two-frame look-ahead and a rolling KV cache. The clever training idea: how do you teach 'good listening' without labels? They make a bad example by running the same model with the user's input removed, so it ignores you, and train the model to prefer the version that pays attention. About half a second of delay on an H100, nearly seven times faster than the previous system, and people preferred it over eighty percent of the time. [Play.] The left column is the real person, the middle is the previous method, the right is Avatar Forcing. Watch the right one while the user is talking.

中文讲解：最新的一环是“倾听”。KAIST与南洋理工的Avatar Forcing (CVPR 2026) 把你的音频和头部动作作为输入，你说话时它会点头、微笑或因你的笑话发笑。技术上是扩散强制：每块10帧、前瞻2帧、滚动KV缓存。训练巧思：如何不用标注教会“好的倾听”？用同一模型去掉用户输入（即无视你）生成坏样本，让模型偏好关注你的版本。H100上约半秒延迟，比之前系统快近7倍，超过80%的比较中更受青睐。播放：左列为真人，中间为之前的方法，右列为Avatar Forcing，注意看用户说话时右边的反应。

72. Inside Avatar Forcing

Inside Avatar Forcing. The key is the input on the left: not just the avatar's own audio, but the user's audio and the user's motion, through a dual motion encoder. Then a causal generator writes the next block of motion, conditioned on the previous clean blocks and a reference frame, and a decoder renders it. The listening behaviour comes from that extra input, the other person.

中文讲解：Avatar Forcing内部：关键在左边的输入——不仅有数字人自己的音频，还有用户的音频和动作，经过双运动编码器；然后因果生成器根据之前的干净块和参考帧写出下一块动作，由解码器渲染。倾听行为正来自这个额外输入：另一个人。

73. Real-time scoreboard, 2024–2026

Here's a scoreboard of real-time systems. Faces only, like VASA-1 and Hedra's Livatar, are very fast: Livatar reports a hundred and forty-one frames per second and 0.17 seconds on a mid-range A10 GPU. Systems that generate sound too, like OmniForcing and Hallo-Live, are around three-quarters of a second to one second. Full-body StreamAvatar is about 1.2 seconds on two H800s. And Wan-Streamer from Alibaba, one model for everything, about 200 milliseconds model-side, but only at 192p for now. Careful: these aren't directly comparable. Resolution, face versus body, and what counts as 'delay' all differ between papers. That's another evaluation lesson for next week.

中文讲解：实时系统一览：只做脸的VASA-1和Hedra的Livatar很快，Livatar在中端A10上达到141帧/秒、0.17秒；同时生成声音的OmniForcing和Hallo-Live约0.7到1秒；全身的StreamAvatar在两张H800上约1.2秒；阿里的Wan-Streamer一个模型做所有事，模型端约200毫秒，但目前只有192p。注意不能直接比较：分辨率、脸或全身、“延迟”的定义各论文不同——又是下周评估课的一课。

74. Pipeline or one brain?

Two philosophies. The cascade, which most products use: separate models for turn detection, recognition,

language, speech and avatar. Easy to upgrade one part, but each hand-off adds delay, and the tone of your voice is lost the moment it becomes text. The end-to-end approach: Alibaba's Wan-Streamer, from June, is one transformer that takes in text, audio and video and puts out audio and video, in 160-millisecond units, with one GPU 'thinking' and passing its memory to another GPU 'performing'. It can respond to how you said something, not just what. It's a proof of concept at low resolution, but I think this is where things are going.

中文讲解：两种思路。级联（多数产品）：轮次检测、语音识别、大模型、语音合成、数字人各一个模型，易于升级单个部件，但每次交接增加延迟，你的语气一变成文字就丢失了。端到端：阿里6月的Wan-Streamer是一个transformer，输入文本、音频、视频，输出音频和视频，以160毫秒为单位，一张卡“思考”并把记忆传给另一张卡“表演”。它能回应你“怎么说”，而不只是“说了什么”。目前只是低分辨率的概念验证，但我认为这是方向。

75. Wan-Streamer: one model hears, thinks and answers

Here is Wan-Streamer in a live call. [Play part.] The person in the small window talks; the character answers with voice and video, from a single model. Notice the interruptions and the timing. The resolution is low; this is version 0.2 of a research system. But compare the rhythm of this conversation with the cascade we costed out a few slides ago.

中文讲解：Wan-Streamer的实时通话演示：小窗口里的人说话，角色用声音和视频回应，全部来自一个模型。注意打断和节奏。分辨率较低，这只是研究系统的0.2版，但请把这段对话的节奏与几页前计算过的级联管线对比。

76. From our lab: TalkVid

Let me show you some work from our own lab, because story three runs on data. TalkVid is a talking-head dataset we built with collaborators at CUHK-Shenzhen and Sun Yat-sen University; it's in CVPR 2026 Findings. About twelve hundred and forty hours of talking video, filtered down from over six thousand hours, with seven thousand seven hundred speakers, from children to over sixty, in fifteen languages, mostly at 1080p or higher, and including the upper body, not just the face. And a benchmark, TalkVid-Bench: five hundred clips balanced by age, gender, ethnicity and language, so you can see who a model works for and who it fails.

中文讲解：介绍我们实验室的工作，因为第三个故事靠数据驱动。TalkVid是我们与港中大（深圳）和中山大学合作者构建的说话人数据集，收录于CVPR 2026 Findings：从6000多小时网络视频中筛出约1244小时，7729位说话人，年龄从儿童到60岁以上，15种语言，大部分为1080p以上，并包含上半身。另有TalkVid-Bench基准：500个片段按年龄、性别、族裔和语言平衡，用来检验模型对谁有效、对谁失效。

77. How TalkVid was filtered: 6,000+ hours → 1,244

How do you turn web video into training data? A filtering pipeline. Collect over thirty thousand videos at 1080p or higher. Cut them into shots and drop anything shorter than five seconds or without speech. Score them for visual quality with an aesthetic model and keep only the sharp, well-lit ones. Track motion and drop clips that are jittery, or frozen like a still image. Finally check the head: pose, rotation, is the face complete in frame. We checked the filters against human judges, and they agree about ninety-five percent of the time. The payoff: a model trained on our balanced 160-hour subset gets better video quality scores than the same model trained on older datasets.

中文讲解：如何把网络视频变成训练数据？一套过滤流程：收集3万多个1080p以上视频；切镜头，去掉短于5秒或无语音的片段；用美学模型打分，只保留清晰、光线好的；追踪运动，去掉抖动或静止的片段；最后检查头部姿态和完整性。我们用人工评审核对过滤器，一致率约95%。效果：在我们平衡的160小时子集上训练的模型，视频质量指标优于用旧数据集训练的同一模型。

78. The real filter funnel, and what it keeps

Here's the actual funnel from the paper, with real hours at each stage: from the raw pool on the left, through segmentation, the speech filter, the quality filter, the motion filter and the head filter, down to 1,244 hours. Most data is thrown away, and that's normal. Good datasets are mostly decisions about what not to keep.

中文讲解：这是论文中的真实漏斗，每个阶段都标有小时数：从左侧原始数据，经过切分、语音过滤、质量过滤、运动过

滤和头部过滤，最终剩下1244小时。大部分数据被丢弃，这很正常。好的数据集主要是关于“不保留什么”的决定。

79. What the quality filter drops and keeps

And what the quality filter actually does, with example frames from the paper. Left: dropped, dim and soft footage. Right: kept, sharp, well lit, stable. This matters for artists too: a model learns the look of its data, so a filter like this quietly decides the look of every avatar trained on it.

中文讲解：质量过滤器实际做了什么（论文中的示例帧）：左边被丢弃——昏暗、模糊；右边被保留——清晰、光线好、构图稳定。这对艺术家也重要：模型学会的是数据的“样子”，所以这样的过滤器悄悄决定了所有用它训练的数字人的视觉风格。

80. What runs where (for your projects)

Practically, for your projects: what runs where? On a laptop or a mid-range GPU, MuseTalk does real-time lip-sync by repainting the mouth in a single step, and LatentSync works offline with about eight gigabytes. On one RTX 4090 or 5090 you can run light real-time talking heads, and the big fourteen-billion-parameter avatar models offline at 480p, but not live. Live full-body avatars with sound still need several data-centre GPUs. Getting those to run live on a single consumer card is an open research problem; it's one our lab is working on, and it needs all three tricks at once.

中文讲解：实用参考：在笔记本或中端显卡上，MuseTalk一步重绘嘴部可实时对口型，LatentSync离线约需8GB显存；一张RTX 4090或5090可以跑轻量实时说话人，140亿参数级数字人模型可离线生成480p但不能实时；带身体和声音的实时数字人仍需多张数据中心GPU。让它们在单张消费级显卡上实时运行是开放问题，也是我们实验室的研究方向，需要三个技巧同时使用。

81. Where live characters already are

Where are live characters already being used? In Japan, the Korean company ESTsoft signed an agreement with NTT and the taxi company Nihon Kotsu in February to put its real-time avatar in taxis, interpreting for tourists. ElevenLabs, the voice company, added avatars in June. Character.AI, whose research we saw tonight, is building video calls with fictional characters. And in performance, ABBA Voyage showed digital performers can fill an arena; a responsive version is the obvious next step.

中文讲解：实时角色已经在哪里使用？韩国ESTsoft今年2月与NTT和日本交通签约，把实时数字人放进出租车为游客翻译；语音公司ElevenLabs 6月加入数字人功能；Character.AI正在开发与虚构角色视频通话；表演领域，ABBA Voyage证明数字表演者能填满一个场馆，可互动的版本是下一步。

82. Artists working with their own likeness

For the artists: two works by Holly Herndon and Mat Dryhurst that turn this technology inward. At the 2024 Whitney Biennial, xhairymutantx: they trained a model on costumed images of Herndon, so whatever you prompt, you get a distorted version of her, a likeness she controls. And The Call at the Serpentine in London: they recorded fifteen UK choirs, built singing models with Ircam, and created a data trust so the singers share control over how their voices are used. For your own projects, this is a good pattern: use your own voice and face, or people who said yes, and make the consent visible in the work.

中文讲解：艺术家案例：霍莉·赫恩登与马特·德赖赫斯特的两件作品。2024年惠特尼双年展的《xhairymutantx》用赫恩登的变装照片训练模型，无论输入什么都生成变形的“她”——一个由艺术家自己控制的肖像。伦敦蛇形画廊的《The Call》录制了15个英国合唱团，与Ircam合作构建歌唱模型，并建立数据信托让合唱团员共同决定声音的用途。你的项目可以借鉴：用自己或同意者的声音和面孔，并让授权成为作品的一部分。

83. Would you talk to a live avatar of your teacher? Of someone who has died?

For the room. Would you talk to a live avatar of your teacher, say, for office hours at 3 a.m.? And harder: would you talk to a live avatar of someone who has died, built from their videos? What would make it feel respectful, and what would make it feel wrong? Two minutes with your neighbour.

中文讲解：讨论：你会和老师的实时数字人聊天吗？比如凌晨3点的答疑。更难的问题：你会和一位已故者的实时数字人对话吗？什么会让它显得尊重，什么会让它显得不对？和邻座讨论两分钟。

84. Who is the performer?

Story four. We've seen that a model can now give a performance, in real time, with a voice. So who is the performer? The person whose face it is? The actors whose work trained it? The person who prompted it? Or the character itself? This year, that stopped being a seminar question. It became a contract, several laws, and a feature film.

中文讲解：第四个故事。模型现在能实时、带声音地表演，那么谁是表演者？面孔的主人？训练数据中的演员？写提示的人？还是角色本身？今年这不再只是课堂问题，而成了合同、法律和一部长片。

85. Tilly Norwood: an “actress” who never existed

Meet Tilly Norwood. She was created by the London AI studio Particle6, founded by the actor and comedian Eline Van der Velden. In July 2025 she starred in a comedy sketch called ‘AI Commissioner’. In September, at the Zurich Film Festival, her creator said talent agencies were interested in representing her, and the reaction was furious: actors including Emily Blunt and Melissa Barrera objected, and big agencies said they would not sign her. In July this year Particle6 announced a feature film, *Misaligned*, with Norwood in the lead. And last month she did a press tour of seventy-five interviews, including one on Piers Morgan's show where she switched into Cantonese for fifteen seconds. That clip has been viewed over twelve million times.

中文讲解：认识一下Tilly Norwood。她由演员兼喜剧人艾琳·范德维尔登创立的伦敦AI工作室Particle6创造。2025年7月出演喜剧短片《AI Commissioner》；9月苏黎世电影节上创作者称有经纪公司感兴趣，引发强烈反弹，艾米莉·布朗特、梅丽莎·巴雷拉等演员反对，大经纪公司表示不会签约。今年7月Particle6宣布由她主演长片《Misaligned》。上个月她做了75场采访的宣传，在皮尔斯·摩根的节目上突然说了15秒粤语，该片段播放量超过1200万。

86. Watch: synthetic performers, from sketch to stage

Let's watch a little. [Play about thirty seconds of the AI Commissioner sketch.] That's how Tilly Norwood was introduced: in a comedy sketch, as a joke about the industry. In March she released a music video answering the backlash, with eighteen human crew credited and the vocals generated by Suno. Compare that with ABBA Voyage, where the performers themselves agreed and a live band plays, and with Holly Herndon's Holly+, her own voice model that she lets other people use, on her terms. Same technology, four different answers to ‘who agreed?’

中文讲解：看一小段：（播放约30秒《AI Commissioner》）Tilly Norwood就是以一個諷刺行業的喜劇短片亮相的。今年3月她发布了回应争议的音乐录像，署名18位人类工作人员，人声由Suno生成。对比ABBA Voyage（表演者本人同意，现场乐队演奏）和霍莉·赫恩登的Holly+（她自己的声音模型，按她的条件供他人使用）。同样的技术，对“谁同意了”有四种不同回答。

87. Two positions, in their own terms

Here are the two positions, as fairly as I can put them. The actors' union, SAG-AFTRA: Norwood is not an actor but a character generated by a program, trained on the work of countless performers without permission or compensation, with no life experience to draw on; and producers who want to use a synthetic performer must notify and bargain with the union. Her creator's position: she is a creative work, a piece of art, like an animated character, made with a new tool alongside animation, puppetry and CGI, not a replacement for anyone. Keep both in mind; you'll argue one of them in fifteen minutes.

中文讲解：尽量公正地呈现双方立场。演员工会SAG-AFTRA：Norwood不是演员，而是程序生成的角色，在未经许可和补偿的情况下用无数演员的作品训练，没有生活经验可借鉴；制片方使用合成表演者必须通知工会并协商。创作者立场：她是一件创作作品、一件艺术品，就像动画角色，是与动画、木偶、CGI并列的新工具，并不取代任何人。记住两方观点，十五分钟后你们要为其中一方辩论。

88. What the actors won: the 2026 SAG-AFTRA contract

In June this year, SAG-AFTRA members ratified a new four-year contract with the studios, with over ninety-one percent voting yes. On synthetic performers: producers say they won't use one in a human role unless it adds ‘significant additional value’, and they have to go to the union first. The chief negotiator was

honest that this is a strong disincentive, not a ban. On digital replicas, copies of real actors: informed consent and fair pay, including for actors who have died; a real business reason before anyone is scanned; no replica can replace a performer during a strike; and extra protection for minors. Video-game voice and motion actors won similar consent rules after an eleven-month strike that ended in July 2025.

中文讲解：今年6月，SAG-AFTRA成员以超过91%的赞成票批准了与制片公司的四年新合同。关于合成表演者：制片方承诺除非带来“显著附加价值”，否则不在人类角色中使用，并须先与工会沟通——首席谈判代表坦言这是强力约束而非禁令。关于数字复制体：需知情同意和公平报酬（包括已故演员）；扫描前需正当商业理由；罢工期间不得用复制体替代演员；对未成年人有额外保护。游戏配音与动捕演员在2025年7月结束的11个月罢工后也获得了类似的同意规则。

89. The law, region by region (October 2026)

The law is catching up, unevenly. California, since January 2025: a contract clause that lets a studio make your digital replica is void if it's vague and you had no lawyer or union, and replicas of dead performers need the estate's consent. In the US federally, the NO FAKES Act would create a right to your own voice and likeness; it passed committee in June but was blocked in the Senate three weeks ago, so it is not law. In mainland China, since September 2025, AI-generated content must carry both a visible label and a hidden metadata label, which applies to anything you publish there. In the EU, since August this year, deepfakes must be disclosed. And Denmark has a bill in parliament this month that would protect your face and voice almost like copyright. I'm not a lawyer, and this isn't legal advice, but the practical rule for your work is simple: label it, and get consent.

中文讲解：法律在追赶，但不均衡。加州自2025年1月起：措辞模糊且演员无律师或工会代表的数字复制条款无效；已故演员的复制体需遗产方同意。美国联邦层面的NO FAKES法案将确立对自己声音和形象的权利，6月通过委员会但三周前在参议院受阻，尚非法律。中国大陆自2025年9月起要求AI生成内容同时带有显式标识和隐式元数据标识。欧盟自今年8月起要求披露深度伪造。丹麦本月有法案进入议会，将像版权一样保护面孔和声音。这不是法律建议，但对作品而言规则很简单：标注，并取得同意。

90. Your voice is part of your identity

Three cases about voice, each with a different lesson. Lehrman and Sage versus Lovo: two voice actors were paid a few hundred dollars on Fiverr for what they were told were research recordings, which became clone voices sold to customers. Last year a New York court let their claims under publicity rights and contract go forward. Lesson: what the recording is for matters. 'Sky': Scarlett Johansson declined to voice ChatGPT, then a voice that many thought sounded like her appeared; OpenAI paused it and said it was another actress's natural voice. Lesson: resemblance alone can be a problem. And Darth Vader in Fortnite: James Earl Jones's estate consented, but the union still complained that an AI voice took paid work away from actors. Lesson: consent of one person doesn't settle the effect on everyone else.

中文讲解：三个关于声音的案例，各有教训。Lehrman与Sage诉Lovo：两位配音演员在Fiverr上被支付几百美元录制所谓“研究用”音频，结果成了商业克隆声音；去年纽约法院允许其公开权和合同诉求继续审理——录音用途很重要。“Sky”：斯嘉丽·约翰逊拒绝为ChatGPT配音，随后出现了许多人认为像她的声音，OpenAI暂停并称是另一位演员的天然声音——仅仅相似就可能是问题。《堡垒之夜》中的达斯·维达：琼斯的遗产方同意了，但工会仍投诉AI声音夺走了演员的工作——一个人的同意不能解决对其他人的影响。

91. What audiences say (NRG × TheWrap, March 2026)

What do audiences think? A survey by the research firm NRG with TheWrap in March: fifty-six percent say AI actors will never be as good as human actors; only seven percent say they already are; and about three in five say releasing a film with an AI-written script is unacceptable. But here's the twist: the largest group said AI use wouldn't change their interest in seeing a film. Audiences judge the work, and also the story of how it was made. That's useful to know when you write your artist statements.

中文讲解：观众怎么看？NRG与TheWrap今年3月的调查：56%认为AI演员永远不如人类；只有7%认为已经一样好；约五分之三认为上映AI编剧的电影不可接受。但有个反转：最大的群体表示AI的使用不会改变他们看电影的兴趣。观众既评判作品，也评判它是如何被制作的——写创作自述时值得记住。

92. Two flashpoints from this year

Two flashpoints from this year with the same kind of tools and opposite receptions. In February, a viral clip made with Seedance 2.0 showed two famous Hollywood actors fighting. Neither had agreed. The Motion Picture Association and SAG-AFTRA condemned it, and ByteDance pledged fixes. In June, at Tribeca, the artist Ash Koosha premiered *Dreams of Violets*, made with Kling, Claude and Gemini, about this year's protests in Tehran, billed as the first fully AI-generated feature at a major festival. The difference wasn't the model. It was whose likeness, whose story, and who agreed.

中文讲解：今年两个工具相似但反响相反的事件。2月，一段用Seedance 2.0制作的爆红片段展示两位好莱坞明星打斗，两人都未同意；美国电影协会和SAG-AFTRA予以谴责，字节承诺整改。6月翠贝卡电影节上，艺术家阿什·库沙首映了用可灵、Claude和Gemini制作、关于今年德黑兰抗议的《Dreams of Violets》，被称为首部在主要电影节展映的全AI长片。区别不在模型，而在于用了谁的形象、讲谁的故事、谁同意了。

93. A checklist for your Assignments 2 and 3

So here's a checklist for Assignments 2 and 3, both due November 30. Consent: only clone or animate a real person's face or voice if they agreed, in writing; using yourself is the easiest. Label: say clearly which faces, voices and sounds are generated; in mainland China and the EU that's now required. Credit: name the models, the voice references, and every human who performed, recorded or wrote. And keep your sources, last week's rule: prompts, seeds, reference audio, model versions and raw takes.

中文讲解：作业二、三（均为11月30日截止）的清单：同意——只有对方书面同意才克隆或驱动真人的面孔或声音，用自己最简单；标注——说明哪些面孔、声音和音效是生成的，在中国大陆和欧盟这已是法定要求；署名——写明所用模型、声音参考，以及每位参与表演、录音或写作的人；保留素材——上周的规则：提示词、种子、参考音频、模型版本和原始素材。

94. Debate: “An AI performer that never existed should be allowed to win an acting award.”

Debate time. The motion: ‘An AI performer that never existed should be allowed to win an acting award.’ Left side of the room argues for, right side against. Two minutes to prepare in groups, then two speakers per side, one minute each, and the whole room votes. Use tonight's material: synchresis, OmniHuman's acting plan, the union's position, the creator's position, and the audience survey.

中文讲解：辩论时间。辩题：“一个从未存在的AI表演者应当有资格获得表演奖。”教室左边正方，右边反方。小组准备两分钟，每方两位发言人各一分钟，最后全班举手投票。请用上今晚的材料：合成同步、OmniHuman的表演计划、工会立场、创作者立场和观众调查。

95. Arguments to start from

If you're stuck, here are starting points, but try to find better ones. For: animated performances already move us, so awards could judge the result, not the method; or the award could go to the humans who made the choices; or there could be a separate category. Against: acting awards honour a person's lived skill and the risk of being seen; the synthetic performance is averaged from thousands of uncredited actors; and prizes would reward replacing the very people the model learned from.

中文讲解：如果没思路，可以从这些论点出发，但尽量找到更好的。正方：动画表演早已打动我们，奖项可以评判结果而非方法；奖项可以颁给做出选择的人类；或另设奖项类别。反方：表演奖表彰的是一个人的亲身技艺和在镜头前的冒险；合成表演是无数未署名演员的平均；颁奖等于奖励取代模型学习对象的人。

96. Four stories, one line each

Four stories, one line each. Sound: picture and sound now share one model and one clock, and the human job is choosing what to hear. Performance: direct with references and playable actions, and check the p, b and m sounds. Real time: chunks, fewer steps and fewer bits turn a renderer into a conversation partner that listens. And the performer: consent, credit and labels are now contracts and law, not just good manners.

中文讲解：四个故事，各一句话：声音——画面和声音共享一个模型、一个时钟，人的工作是选择听什么；表演——用参考和可表演的动作来导演，检查p、b、m；实时——分块、少步、低比特把渲染器变成会倾听的对话者；表演者——同意、署名和标注如今是合同和法律，而不只是礼貌。

97. Tonight → your assignments (both due Nov 30)

How does tonight connect to your assignments? For Assignment 2, when you reverse-engineer an AI video, reverse-engineer the sound as well: which layers were generated, and by which model? Then remix it, with a native-audio model or your own voice track. For Assignment 3, your game: try giving one character a voice and a face that reacts. A talking portrait with MuseTalk or LatentSync, or a short generated cut-scene with dialogue, is enough. Both are due November 30. And use your own voice, or a friend's who agreed, and label what's generated.

中文讲解：与作业的联系：作业二在逆向分析AI视频时，也要逆向分析声音——哪些层是生成的、用的什么模型？然后用原生音频模型或你自己的配音来重混。作业三的游戏：试着给一个角色加上声音和会反应的面孔，用MuseTalk或LatentSync做会说话的肖像，或一段带对白的生成过场动画即可。两者均11月30日截止。请用自己的声音或获得书面同意的朋友的声音，并标注生成内容。

98. Three readings for this week

Three readings for this week, one for each kind of student, and you reflect on any one for the quiz.

Technique: the LTX-2 technical report. Read the abstract, look at the architecture figure we discussed, and listen to two of the demos. Art and tech: The Call, by Holly Herndon and Mat Dryhurst at the Serpentine: choirs, singing models, and a data trust. Culture: Variety's report on Tilly Norwood's first feature film, read together with the union's statement. And two optional technical papers: Avatar Forcing, and our TalkVid dataset paper.

中文讲解：本周三篇阅读，测验时任选一篇反思。技术：LTX-2技术报告，读摘要、看我们讨论过的架构图、听两个演示。艺术与技术：霍莉·赫恩登与马特·德赖赫斯特在蛇形画廊的《The Call》——合唱团、歌唱模型和数据信托。文化：Variety关于Tilly Norwood首部长片的报道，配合工会声明阅读。另有两篇选读技术论文：Avatar Forcing和我们的TalkVid数据集论文。

99. Quiz 08 · five short questions

Quiz 08 is online now: two technique questions, on joint audio-video generation and on the real-time tricks; two art questions, on synchresis and on directing a line; and one reading reflection. One to three sentences each. As always, it's due at the end of the semester, and you can edit your answers until then.

中文讲解：第八次测验已上线：两道技术题（联合音视频生成、实时技巧），两道艺术题（合成同步、台词导演），一道阅读反思。每题一到三句话。照例学期末截止，之前可随时修改。

100. Every demo from tonight, in one place

This slide collects every demo from tonight so you can find them again. They're also on the course website, on the Lecture 8 page under Demos. All the videos belong to their creators and were shown for teaching.

中文讲解：本页汇总今晚所有演示，课后也可在课程网站第八讲页面“Demos”处找到。所有视频版权归原作者，仅用于教学。

101. Next week: critique, evaluation, provenance

Next week, Week 9: critique, evaluation and responsible release. We'll generalise tonight's trap, a model 'beating' real video on a metric, and ask what our scores really measure. Please bring one generated clip with sound for a studio critique. And we'll cover model cards, watermarks and labels.

中文讲解：下周第九讲：评价、评估与负责任发布。我们会把今晚的陷阱（模型在指标上“超过”真实视频）推广开来，讨论分数到底衡量什么。请带一段带声音的生成片段参加工作室点评。我们还会讲模型卡、水印和标注。

102. Sound, performance, presence

To sum up tonight in one line: generated video learned to speak, to act, and to listen. Our job is to choose what we hear, to direct how it's said, and to ask whose voice it is. Thank you, and see you next week.

中文讲解：一句话总结：生成视频学会了说话、表演和倾听。我们的工作选择听什么、导演怎么说，并追问这是谁的声音。谢谢，下周见。

103. Do it yourself

The last few slides are an appendix for after class: practical tools, a workflow, prompts, and checks you can use for Assignments 2 and 3. I'll go through them quickly; they're all on the website.

中文讲解：最后几页是课后用的附录：实用工具、工作流程、提示词和检查清单，可用于作业二和作业三。我快速过一遍，网站上都有。

104. Toolkit: open models for each job

Here's a toolkit table: for each job, an open model, its size or hardware as reported by the authors, and the licence. Picture and sound together: LTX-2, or the lighter Kandinsky 6 Lite and NAVA. Two-person dialogue: Ovi or MultiTalk. Adding Foley to a silent clip: MMAudio or Tencent's HunyuanVideo-Foley; note MMAudio's weights are non-commercial. Lip-syncing existing footage: LatentSync or MuseTalk, which run on modest GPUs. Re-dubbing: InfiniteTalk. Talking or singing bodies: Wan2.2-S2V or HunyuanVideo-Avatar. Always check the licence before commercial use.

中文讲解：工具表：每项任务对应的开源模型、作者报告的规模或硬件需求及许可证。画面加声音：LTX-2，或更轻的Kandinsky 6 Lite和NAVA；双人对话：Ovi或MultiTalk；给无声片段加拟音：MMAudio或腾讯HunyuanVideo-Foley（注意MMAudio权重仅限非商业）；给现有素材对口型：LatentSync或MuseTalk，中等显卡即可；整段重配音：InfiniteTalk；说话或唱歌的身体：Wan2.2-S2V或HunyuanVideo-Avatar。商用前务必核对许可证。

105. Workflow: a talking character in one evening

A workflow you could do in one evening. One: record the line yourself on your phone in a quiet room. Your timing is the performance. Two: generate or paint one front-facing image of the character with the mouth clearly visible. Three: animate it with an audio-to-video model, or lip-sync an existing clip. Four: add ambience and Foley under the voice. Five: run the checklist on the next slide and re-generate only the shot that fails. Using your own voice avoids every consent problem from story four.

中文讲解：一个晚上可以完成的流程：一、用手机在安静房间自己录台词，你的节奏就是表演；二、生成或绘制一张正面、嘴部清晰、光线好的角色图；三、用音频到视频模型驱动，或给现有片段对口型；四、在人声下加入环境声和拟音；五、按下一页清单检查，只重新生成不合格的镜头。用自己的声音可以避开第四个故事中的所有授权问题。

106. Review checklist for a generated performance

And the review checklist. Lips: full closure on p, b and m. Timing: scrub frame by frame at a hard consonant or an impact and check the sound lands on that frame. Voice: the same voice across shots, breaths in human places. Eyes and body: motivated gaze, hands in rhythm. Room: reverb that matches the space, ambience continuous across cuts. And credits and labels written down before you submit.

中文讲解：检查清单：嘴——p、b、m完全闭合，f、v牙齿碰唇；时机——在硬辅音或撞击处逐帧拖动，声音是否落在那一帧；声音——各镜头声音一致，呼吸和停顿在合理位置；眼睛与身体——视线有动机，手部节奏与乐器一致；空间——混响与空间相符，环境声跨剪辑连续；署名与标注——提交前写清模型、声音参考、授权和AI标注。

107. Measure sync yourself (two commands)

Finally, how to measure sync on your own work. The reference SyncNet code is on GitHub; three commands give you the offset in frames, Sync-D and Sync-C. And to train your own eyes, use ffmpeg to make a copy with the sound delayed by 125 milliseconds, the edge of what viewers tolerate, and compare. That's exactly how I made tonight's lag game.

中文讲解：最后，如何测量你自己作品的同步：GitHub上的参考SyncNet代码，三条命令即可得到偏移帧数、Sync-D和Sync-C。训练眼睛的方法：用ffmpeg生成一个声音延迟125毫秒（观众容忍的边缘）的副本，并排对比。今晚的延迟游戏就是这样做的。