

AMCC 5160 · Lecture 08 — Reading guide

All three connect directly to tonight's stories. Two of them come with videos or sound, so listen as well as read. For Quiz 08, reflect on any one of the three. Two optional technical readings are listed last.

Reading 1 / Technique: LTX-2: Efficient Joint Audio-Visual Foundation Model

Lightricks · arXiv 2601.03233 | January 2026 · <https://arxiv.org/abs/2601.03233>

What to do: Read the abstract and the introduction, and study the architecture figure (the two streams). Then listen to two demos on the GitHub page. You do not need the training details.

LTX-2 generates video and sound together with two transformer streams: a large one for picture (about 14B parameters) and a smaller one for sound (about 5B), which exchange information in every block through cross-attention that uses only time positions. It makes up to 20 seconds of video with stereo audio, and its weights are open.

Questions to keep in mind:

- Why is the audio stream so much smaller than the video stream?
- What does it mean that the two streams share “only the time positions” when they attend to each other?
- Listen to one demo: which sounds feel generated together with the picture, and which feel pasted on?

Story 1: “Inside LTX-2: two streams, one clock” .

Code and weights: <https://github.com/Lightricks/LTX-2>

Reading 2 / Art + tech: Holly Herndon & Mat Dryhurst: The Call

Serpentine Galleries, London · Exhibition | October 2024 – February 2025 ·

<https://www.serpentinegalleries.org/whats-on/holly-herndon-mat-dryhurst-the-call/>

What to do: Read the exhibition page and listen to any recordings or videos linked from it. Then look at the Whitney page for their earlier work *xhairymutantx* (2024).

Herndon and Dryhurst recorded fifteen choirs across the UK, built singing models with Ircam and others, and set up a data trust so that the choristers share control over how their voices are used. It is an artwork about AI voices that makes consent and shared ownership part of the work.

Questions to keep in mind:

- What does the data trust change about who owns a voice model?
- Would you contribute your voice to a project like this? Under what conditions?
- How could you make consent visible in your own Assignment 2 or 3?

Story 3: “Artists working with their own likeness” , and the Story 4 checklist.

Related: *xhairymutantx*, Whitney Biennial 2024 — <https://whitney.org/exhibitions/xhairymutantx>

Reading 3 / Culture: Tilly Norwood to Lead New Movie ‘Misaligned,’ Marking Feature Debut for AI ‘Actor’

Alex Ritman, Variety · July 2026 · <https://au.variety.com/2026/film/news/ai-actor-tilly-norwood-movie-debut-misaligned-38324/>

What to do: Read the article (about 5 minutes), then read SAG-AFTRA's September 2025 statement on Tilly Norwood (linked below) and compare the two positions.

Particle6, the London studio that created the AI character Tilly Norwood, announced a comedy-drama feature, *Misaligned*, with Norwood in the lead: an AI with no body or childhood who drops her guardrails. Particle6 calls it a hybrid production with human directors, writers and editors alongside AI specialists. The

actors' union says Norwood is not an actor but a character generated from performers' work without permission.

Questions to keep in mind:

- Which claims in the article are facts, and which are promotion?
- Is Tilly Norwood closer to an animated character or to a digital replica? Why?
- Who should be credited for her performance?

Story 4 and tonight's debate.

SAG-AFTRA statement, as reported by ICG: <https://www.icg600.com/news-articles/news/sag-aftra-slams-ai-actress-tilly-norwood-not-actor-it-has-no-life-experience-draw-no-emotion>

Optional / Technique: Avatar Forcing: Real-Time Interactive Head Avatar Generation for Natural Conversation

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, Sung Ju Hwang · KAIST, NTU Singapore, DeepAuto.ai · CVPR 2026 | arXiv 2601.00664 · <https://arxiv.org/abs/2601.00664>

What to do: Optional. Read the abstract and watch the listening examples on the project page. Not needed for Quiz 08.

An avatar that reacts while you talk: it takes your audio and head motion as input and generates its own motion causally with diffusion forcing, at about 500 ms latency on one H100. It learns good listening without labels by preferring its own output over a version that ignores the user.

Questions to keep in mind:

- Why does listening matter as much as speaking?
- How does the model learn “good reactions” without labelled data?

Story 3: “Avatar Forcing: an avatar that listens” .

Project page: <https://taekyungki.github.io/AvatarForcing>

Optional / Data: TalkVid: A Large-Scale Diversified Dataset for Audio-Driven Talking Head Synthesis

Shunian Chen, Hejin Huang, Yexin Liu et al. · CUHK-Shenzhen, SYSU, HKUST · CVPR 2026 Findings | arXiv 2508.13618 · <https://arxiv.org/abs/2508.13618>

What to do: Optional. Read the abstract and look at the filtering pipeline and the TalkVid-Bench figure. Not needed for Quiz 08.

A 1,244-hour talking-head dataset with 7,729 speakers in 15 languages, filtered from over 6,000 hours of web video, plus a 500-clip benchmark balanced by age, gender, ethnicity and language to show who a model works for.

Questions to keep in mind:

- Why does a talking-head dataset need to be balanced?
- Which filtering step would you consider most important, and why?

Story 3: “From our lab: TalkVid” .

Dataset: <https://github.com/FreedomIntelligence/TalkVid>