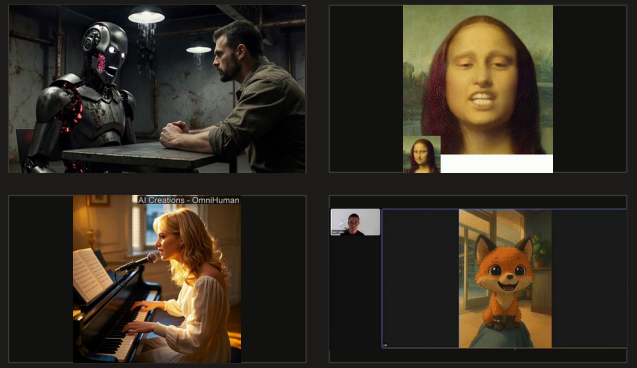
AMCC 5160 · LECTURE 08 · 20 OCTOBER 2026



When AI Video *Learned to Talk*

Sound, performance, and real-time characters · Fall 2026

Harry Yang · Division of Arts and Machine Creativity, HKUST · MIL13002

Sound off: what are they saying?



Play it muted first. Guess the line, the voices and the room · Ovi · Character.AI · click to watch · source ↗

Sound on: one model made all of it



Voices, room tone and timing were generated together with the picture · Ovi · Character.AI · click to watch · source ↗

Four stories from this year

01

Sound is no longer post-production

Picture, voice, effects and music now come out of one model

Veo 3 (May 2025) → Kling 4.0 (Oct 2026)

02

Directing a performance, not a prompt

Lip-sync grew into acting: emotion, timing, gesture, duets

Wav2Lip (2020) → OmniHuman-1.5 (2025)

03

Characters that listen

Real-time avatars that react while you are still talking

Avatar Forcing, CVPR 2026

04

Who is the performer?

Synthetic actors, digital replicas, and the new rules

Tilly Norwood → SAG-AFTRA 2026

Where we've been: seven weeks in one slide

01 · Sep 1

Neural networks to AI filmmaking

02 · Sep 8

Generative editing and art direction

03 · Sep 15

Inside the generator

04 · Sep 22

AI video and art filmmaking

05 · Sep 29

Adaptation and authorship

06 · Oct 6

Quantization for faster inference (Yaofu)

07 · Oct 13

Music, minutes, and machines

08 · Tonight

Sound, performance, real-time characters

Tonight reuses three tools: tokens (Week 3), speed tricks (Weeks 6–7), and authorship (Week 5).

Synchresis: why sound is half the picture



The idea

Sound + image at the same instant → one event (Chion, 1990)



Why Foley works

Fake sounds convince when the timing is right



Why bad lip-sync hurts

Viewers notice sound ~45 ms early or ~125 ms late

Keep this word. It explains both the magic and the failures you will see tonight.

STORY 1 OF 4

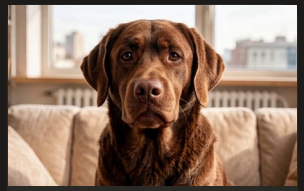
01

Sound is no longer post-production

Picture, voice, effects and music from one model, in one pass



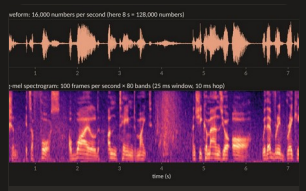
The old pipeline: five tools, five chances to drift



Picture

Silent text-to-video

→



Voice

TTS or a recorded actor

→



Lips

Lip-sync repaints the mouth

→



Foley

Video-to-audio adds effects

→

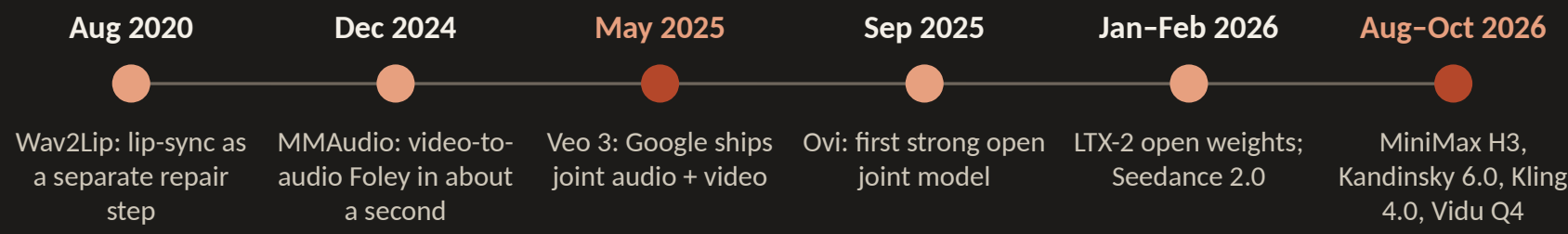


Mix

An editor balances it all

Each hand-off can lose timing, emotion or identity.

Eighteen months, from add-on to native



By mid-2026, almost every frontier video model generates its own sound.

Veo 3: the clip that started it



Voice, wind and sea from one prompt, in one pass · Veo 3 · Google DeepMind · click to watch · source ↗

Source: Google DeepMind, Veo 3 model card (2025)

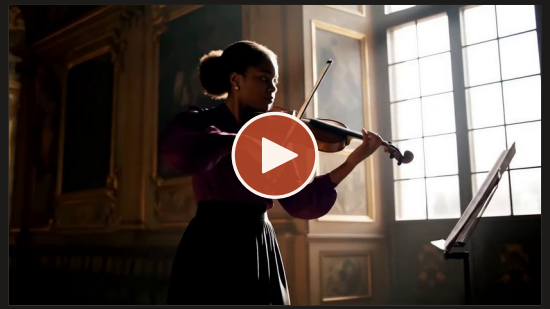
Three kinds of sound, one model



Effects

Sizzle and clatter timed to each toss of the wok.

Veo 3 · Google DeepMind · click to watch · source ↗



Music

The bow and the note move together.

Veo 3 · Google DeepMind · click to watch · source ↗



Dialogue

A detective interrogates a rubber duck. Comedy needs timing.

Veo 3 · Google DeepMind · click to watch · source ↗

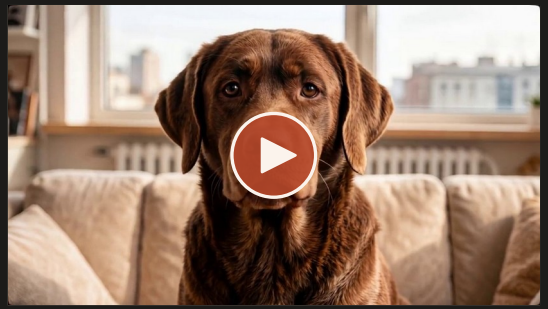
Listen for synchresis: which sound "welds" to the picture best?

Who ships native audio in 2026

Model	Maker	Open?	What stands out
Veo 3.1	Google	Closed	Audio in reference, first/last-frame and Extend modes
Seedance 2.0 / 2.5	ByteDance	Closed	Up to 9 images + 3 videos + 3 audio clips as references
Kling 4.0	Kuaishou	Closed	Up to 30 s per generation (Sep 28 preview)
Vidu Q4 Preview	ShengShu	Closed	Voice references for consistent characters (Oct 7)
LTX-2 / 2.5	Lightricks	Open weights	19B: 14B video + 5B audio stream
Ovi	Character.AI	Open (Apache-2.0)	Twin towers; audio tower trained from scratch
MiniMax H3	MiniMax	Open weights	33B single stream; 32 kHz stereo audio
Kandinsky 6.0	Kandinsky Lab	Open (MIT)	3B Lite and 29B Pro; 44 kHz audio

Sources: official model cards, technical reports and release notes (Google, ByteDance Seed, Lightricks arXiv 2601.03233, Character.AI arXiv 2510.01284, MiniMax, Kandinsky arXiv 2610.05608).

Open weights, this summer and autumn



Kandinsky 6.0 (Oct 2026)

MIT license; 3B Lite and 29B Pro; 44 kHz audio.

Kandinsky 6.0 · Kandinsky Lab · click to watch · source ↗



MiniMax H3 (Aug 2026)

33B single stream; 32 kHz stereo; 11 languages of dialogue.

MiniMax H3 · MiniMax · click to watch · source ↗



NAVA (May 2026)

6.3B, Apache-2.0; best open lip-sync on Verse-Bench.

NAVA · Baidu ERNIE · click to watch · source ↗

You can download and run all three; check each license before commercial use.

Closed models: Seedance 2.0's range



A painting comes alive

Chinese ink-painting style, with hooves and wind generated to match.

Seedance 2.0 · ByteDance Seed · click to watch · source ↗



A concert hall

Hands, music and the room's reverb, together.

Seedance 2.0 · ByteDance Seed · click to watch · source ↗

Under the hood: what is sound to a model?

48,000

Numbers per second per channel: raw audio is a pressure wave, sampled

→

Mel

Many short FFTs turn it into a picture of pitch over time (a spectrogram), warped to how we hear

→

25 / s

An audio VAE squeezes it to about 25 latent vectors per second (LTX-2: 128 numbers each)

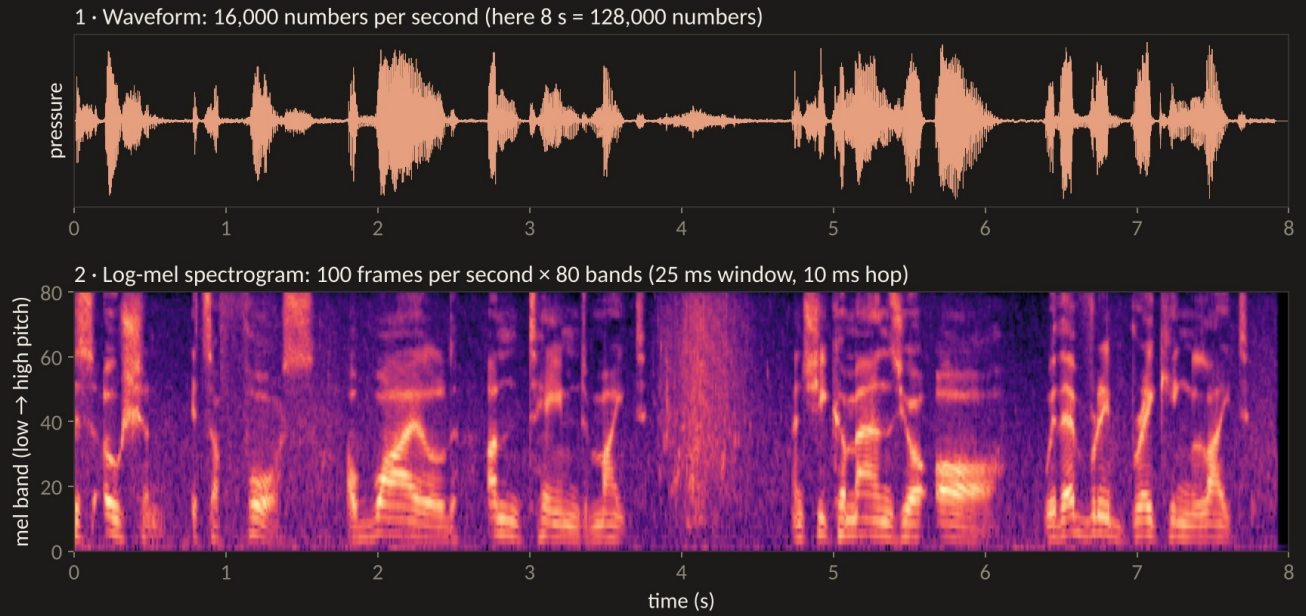
→

Tokens

Now sound is a short sequence the transformer can attend to, like video tokens

Adapted from CMU 10-423 Generative AI, Lecture 24 “Audio Understanding and Synthesis” (Gormley & Nayebi, Fall 2025); numbers from LTX-2 (arXiv 2601.03233).

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



Computed from the demo clip for this lecture. Horizontal stripes are the harmonics of a voiced vowel; vertical bursts are consonants and splashes; the hiss between words is wind and sea. A mel spectrogram is literally an image of sound.

Worked example: 10 seconds of picture vs sound

240

frames: 10 s of 720p at 24 fps

→

61

latent frames after the video VAE (4× in time)

→

~220k

video tokens (3,600 per latent frame, Week 7)

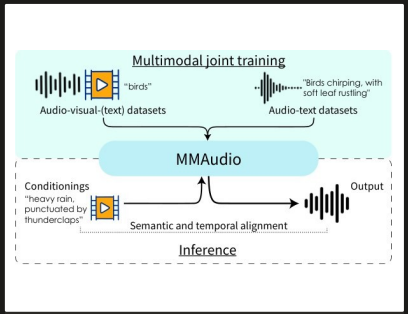
→

250

audio tokens (25 per second, LTX-2)

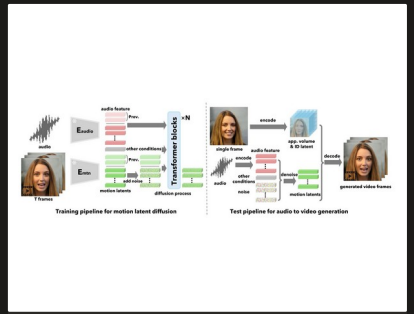
≈ 880× fewer tokens for sound, so LTX-2 gives picture a 14B stream and sound a 5B stream: an “asymmetric dual-stream” model.

Four ways to wire sound and picture



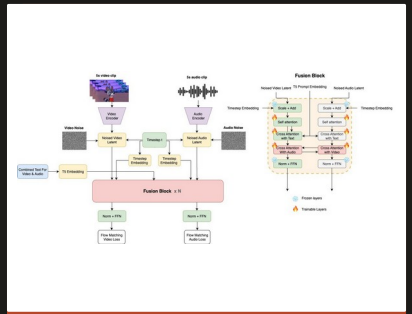
Picture → sound

MMAudio, HunyuanVideo-Foley



Sound → picture

VASA-1, OmniHuman, InfiniteTalk



Two towers that talk

Ovi, LTX-2, Kandinsky 6.0

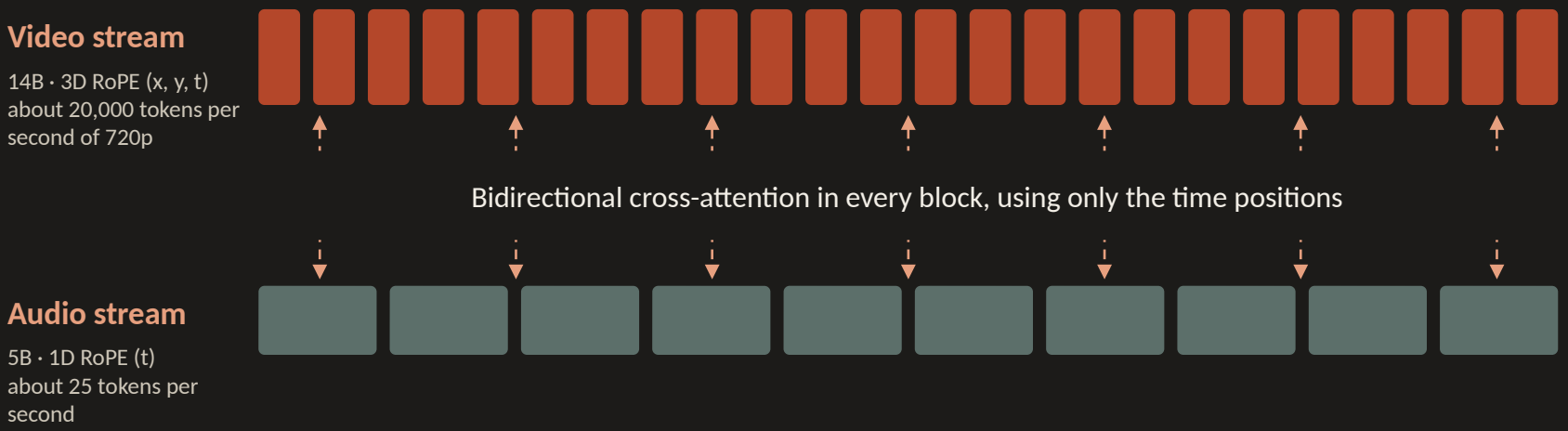


One stream

MiniMax H3, daVinci-MagiHuman

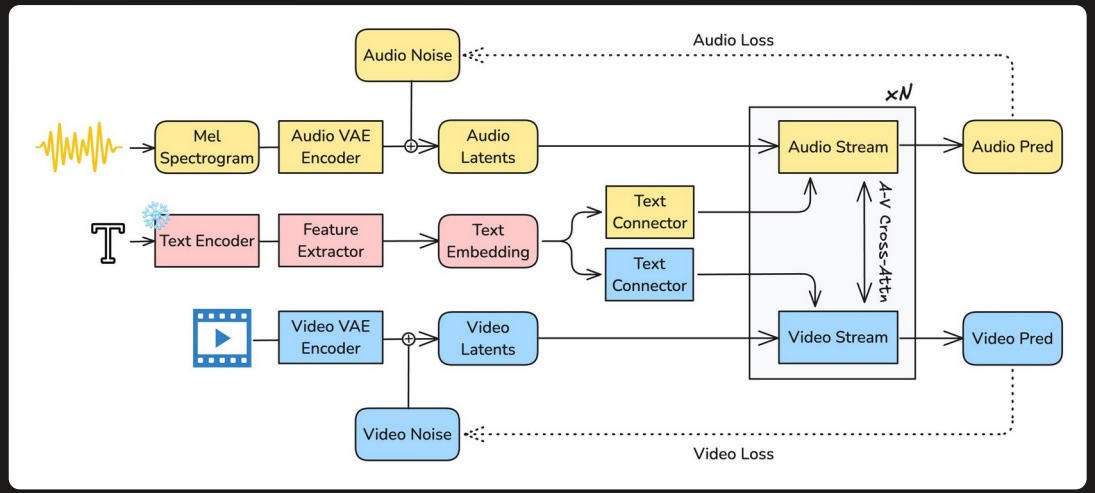
Two towers is the most common open design in 2026.

Inside LTX-2: two streams, one clock



Both streams also share the diffusion timestep. Result: up to 20 s of video with stereo sound; 1.22 s per step on an H100 vs 22.3 s for Wan 2.2-14B. Source: LTX-2, arXiv 2601.03233.

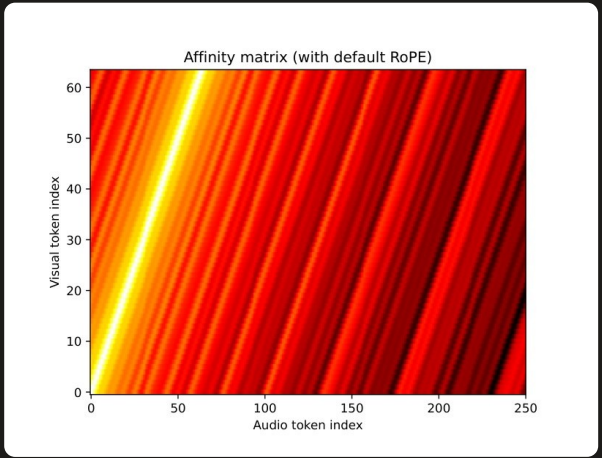
From the technical report: LTX-2 at a glance

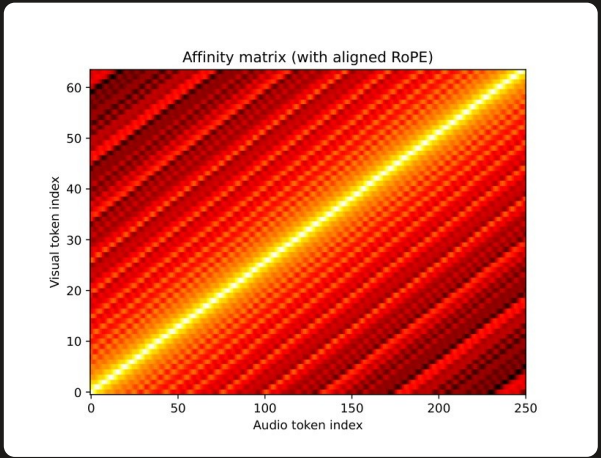


- Two VAEs: one compresses video, one compresses the mel spectrogram
- A text encoder (Gemma 3) feeds both streams through separate “connectors”
- The streams denoise together; the A ↔ V cross-attention is the bridge
- Two losses at the end: one for picture, one for sound

Figure: LTX-2 technical report, Lightricks, arXiv 2601.03233 (2026). Shown for teaching.

Keeping time: why positions must be aligned





Default positions

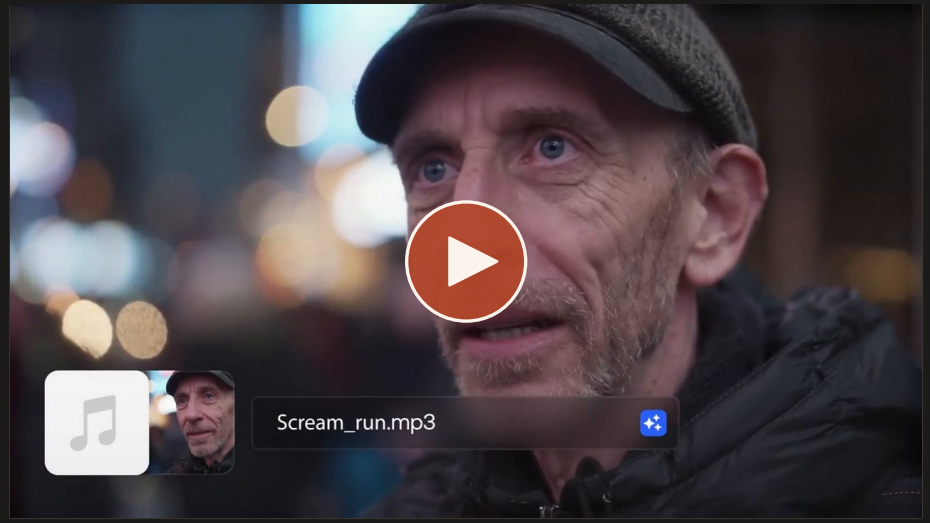
Video and audio tokens arrive at different rates, so “nearest in position” drifts away from “same moment”.

Aligned positions

Scale the positions to seconds and the bright diagonal lines up: each frame attends to its own moment of sound.

Figures: MMAudio (Cheng et al., CVPR 2025), arXiv 2412.15322. LTX-2 and Ovi use the same idea: only time positions are shared across streams.

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



LTX-2 · Lightricks · click to watch

- Audio-to-video: the scream first, the face second
- Open weights; one large GPU
- Also text→AV and image→AV

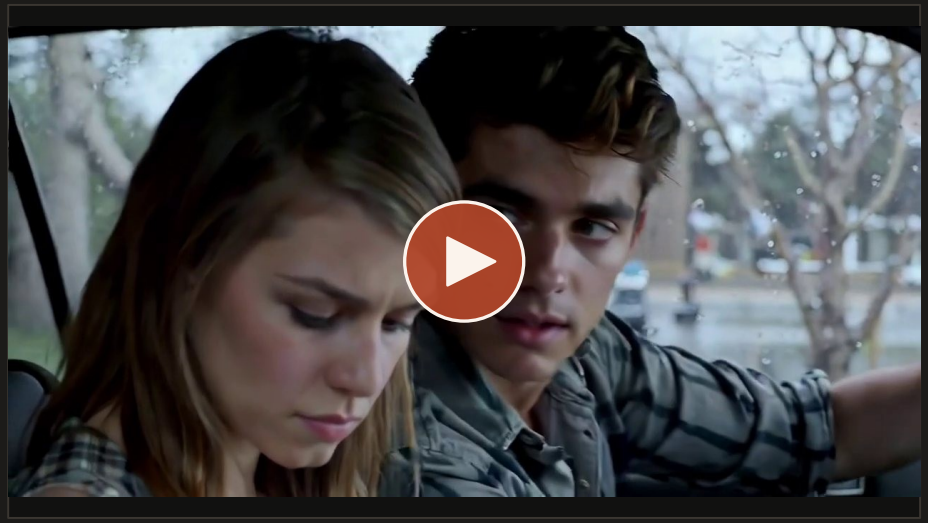
▶ LTX-2 on GitHub

github.com/Lightricks/LTX-2

▶ Technical report (arXiv)

arxiv.org/abs/2601.03233

Ovi: two voices, one take



Ovi · Character.AI · click to watch

- Twin towers: video from Wan 2.2, audio from scratch
- Dialogue written in the prompt
- 5–10 s, 24 fps; Apache-2.0

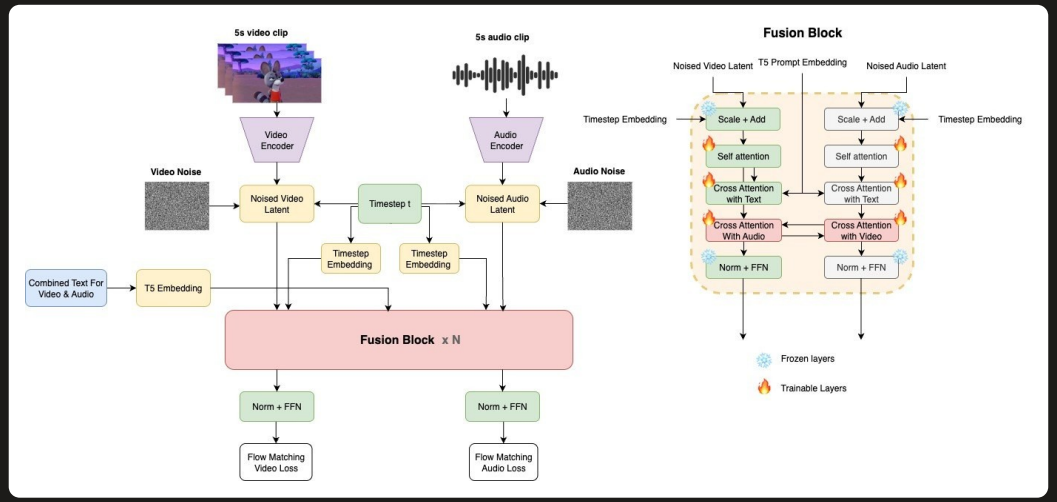
▶ Ovi project page

aaxwaz.github.io/Ovi

▶ Paper (arXiv 2510.01284)

arxiv.org/abs/2510.01284

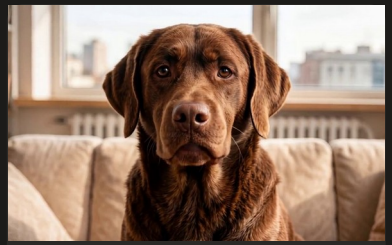
Ovi’s twin towers



- Left tower: video (initialised from Wan 2.2). Right tower: audio (trained from scratch)
- Same shape, block for block, so they can talk at every layer
- Fusion block: self-attention, then cross-attention to text and to the other tower
- Frozen pieces (snowflakes) vs trained pieces (flames): reuse what already works

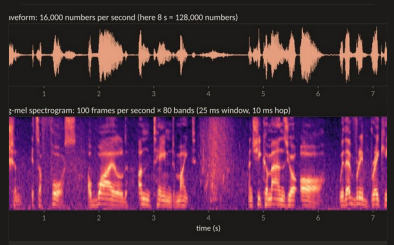
Figure: Ovi, Character.AI and Yale, arXiv 2510.01284 (2025). Shown for teaching.

Recipe: how to teach a video model to hear



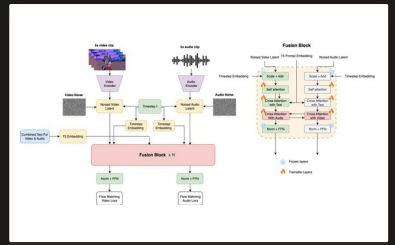
1 · Reuse

Start from a strong video model



2 · Listen

Train audio alone first: ~40M tracks



3 · Join

Train together: ~7M clips with sound



4 · Polish

Fine-tune, RL, distil to 10 steps

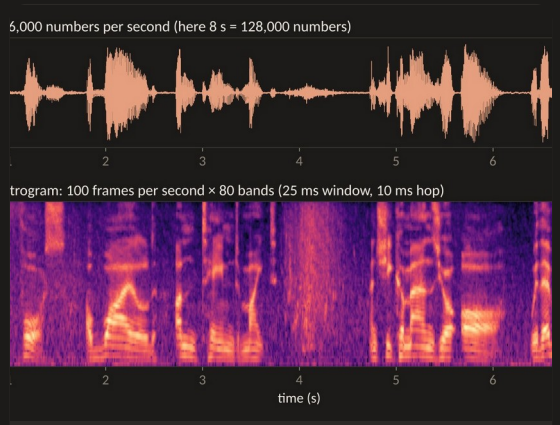
Source: Kandinsky 6.0 Video, arXiv 2610.05608 (Oct 2026). NAVA (Baidu, 6.3B) follows a similar path from Wan 2.2: audio-heavy first, then joint.

Data: finding clips that are really in sync



Throw most of it away

Keep a clip only if a sync score says lips match



Caption the sound too

Voices, effects, music; spoken words in quotes



Know who is talking

One voice fingerprint per speaker (NAVA)

Kandinsky 6.0 kept ~7M segments; NAVA trained on ~15M clips with 160 H100s for ~4 weeks.

Anatomy of an audio-video prompt

Layer	What to write	Example
Picture	Shot, subject, light, camera	Close-up, rainy night, a taxi driver turns to the back seat
Dialogue	Who speaks, the exact words in quotes, how	He says quietly, tired: “You sure this is the address?”
Effects	Sounds tied to visible actions	Wipers squeak once; a turn signal ticks
Ambience	The room or place, always on	Rain on the roof, distant traffic
Music	Inside the scene or score? Mood, level	Low radio jazz from the dashboard, under the voice

Tip from Seedance users: keep a spoken line under about 15 s, and avoid asking for big head turns while a character speaks.

Copy these: three prompts that work

Dialogue scene (Veo, Seedance, Kling, LTX-2, Ovi)

```
Medium close-up, night, inside a taxi in the rain. The driver, 50s, tired, turns to the back seat
and says quietly: "You sure this is the address?" Wipers squeak once. Rain on the roof. No music.
```

Sound-led shot (effects and ambience, no speech)

```
Slow push-in on a wok over a high flame. Oil hisses; each toss of the vegetables lands with a
metallic clatter, in time with the motion. Busy restaurant kitchen ambience in the background.
```

Music inside the scene

```
A street saxophonist plays a slow jazz phrase on a wet city corner at dusk; footsteps and traffic
pass behind him. The music is played on screen; camera holds on his fingers for the last note.
```

Pattern: picture → who says what (in quotes) → effects tied to actions → ambience → music, and say “No music” when you mean it.

Be the Foley artist: what should this sound like?



First muted: list every sound you would add. Then compare with the model · Ovi · Character.AI · click to watch · source ↗

Listen twice: which Foley is better?



Version A

Same burning grassland, sound added by one video-to-audio model.

Clip A · click to watch · source ↗

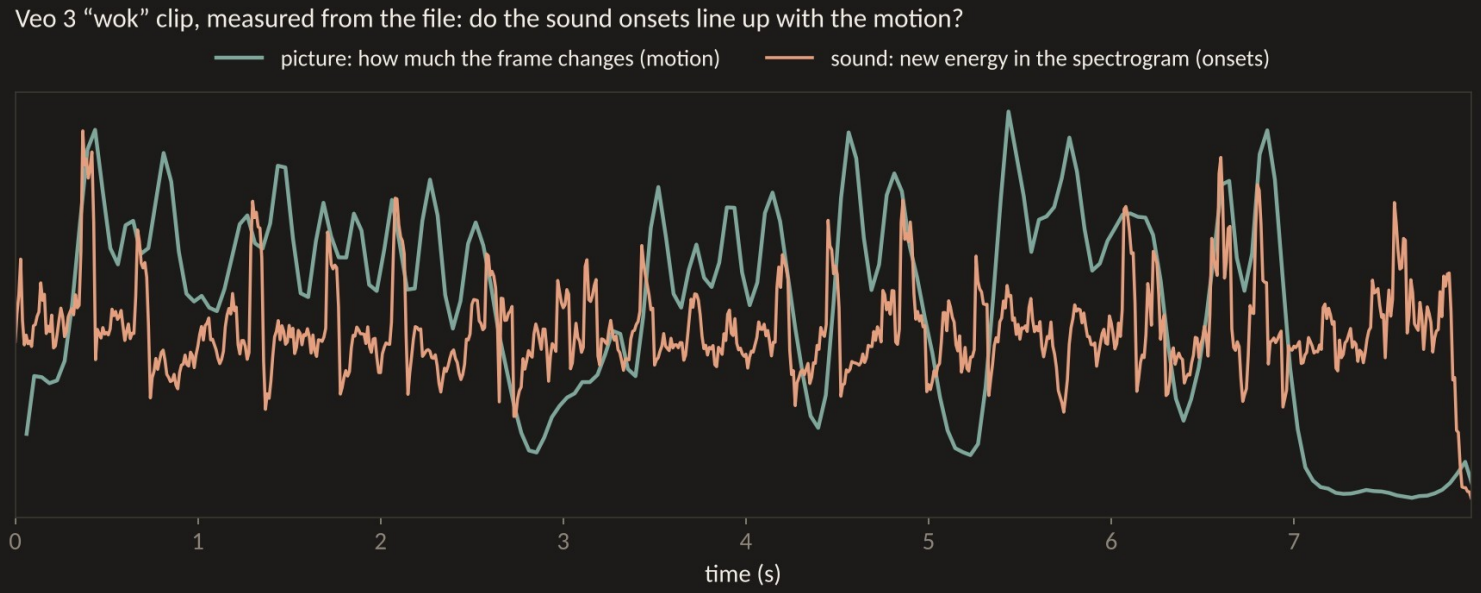


Version B

Same picture, sound from a different model.

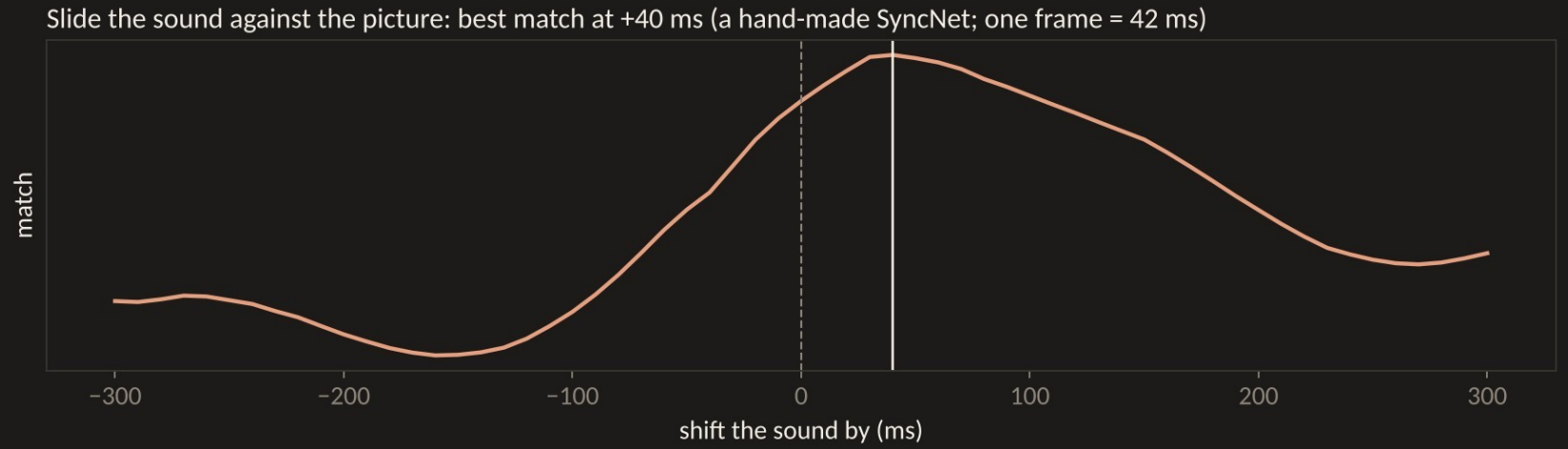
Clip B · click to watch · source ↗

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



Computed from the Veo 3 “wok” clip. Green: how much each video frame changes. Orange: new energy in the sound’s spectrogram (onsets). Many peaks coincide, which is what synchresis needs.

Slide the sound: a hand-made SyncNet



Shift the sound in 10 ms steps and score how well the two curves match. The best match is about one frame (42 ms) from zero, inside the range viewers tolerate (45 ms early to 125 ms late).

Diegetic or not? Music that lives in the scene



Inside the scene (diegetic)

A street saxophonist: the characters could hear this music too.

Ovi · Character.AI · click to watch · source ↗



Played on screen, generated together

UniVerse-1 stitches a video expert and a music expert into one model.

UniVerse-1 · click to watch · source ↗

Where it still breaks



Two people talking

The wrong mouth moves



Motion vs mouth

Big head moves break lip-sync



Audio artefacts

Metallic voices; music over dialogue



Languages

Quality varies by language and dialect



Exact events

The hit lands a few frames early



Same voice, next shot

Fix it with a voice reference

Can you run it yourself? Speed, again

18×

LTX-2 vs Wan 2.2-14B per diffusion step on one H100 (1.22 s vs 22.3 s)

35×

OmniForcing: a 5-s clip in 5.7 s instead of 197 s, by distilling LTX-2 into a streaming model

~25 fps

OmniForcing streams audio + video on one GPU, first chunk in ~0.7 s

Same tricks as Weeks 6–7: fewer steps (distillation), chunks + KV cache, fewer bits. Sources: LTX-2 arXiv 2601.03233; OmniForcing arXiv 2603.11647.

For sound designers and composers: what changes?

What the model now does

- A first mix, already in sync
- Room tone and footsteps for free
- Diegetic music that matches the instrument
- Temp dialogue for pitches

What stays human

- What NOT to hear: silence, focus
- The exact hit, the joke’s timing
- Score: what characters can’t hear
- Whose voice, with consent

FOR THE ROOM

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

And would you credit the model, the prompt writer, or the people whose recordings trained it?

STORY 2 OF 4

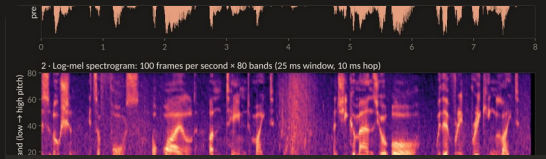
02

Directing a performance, not a prompt

From “do the lips match?” to “is this good acting?”



What is a performance made of?



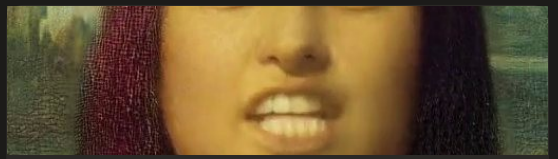
Voice

Pitch, pace, breath, accent



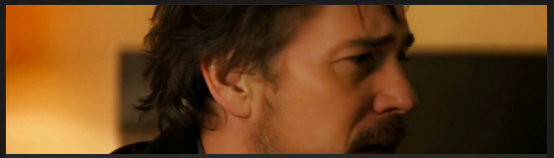
Timing

Pauses, interruptions, the beat



Mouth

Lips, jaw, teeth on every sound



Face and eyes

Brows, blinks, where the gaze goes



Body

Head, shoulders, hands, posture

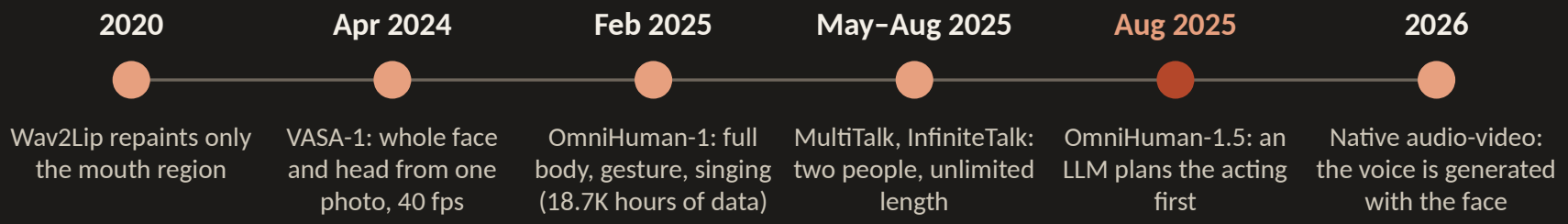


Listening

What you do while others talk

Early models did only the mouth. 2026 models attempt all six; listening is the newest (Story 3).

From lip-sync to acting



Sounds you can see: phonemes → visemes

Sound group	Examples	What the mouth does	Easy to fake?
Bilabials	/p/ /b/ /m/ — “map”, “Bob”	Lips fully close, then release	No: viewers check these
Labiodentals	/f/ /v/ — “five”	Upper teeth touch lower lip	Hard
Rounded vowels	/u/ /o/ — “who”, “go”	Lips push forward into an O	Medium
Open vowels	/a/ — “father”	Jaw drops wide	Easy
Tongue sounds	/t/ /d/ /k/ /g/	Mostly hidden inside the mouth	Easy: invisible

Many sounds share one mouth shape (a viseme). A model must hit the visible ones exactly, especially lip closures.

Under the hood: how a voice drives a face

Voice

16 kHz audio: the line to be spoken

→

Audio encoder

A pretrained speech model (wav2vec 2.0 or Whisper) gives one feature vector per video frame

→

Cross-attention

Each frame's tokens look up "what sound is happening now?" inside the video DiT

→

Frames

Mouth, face and body follow the sound; a reference photo keeps identity

Used by VASA-1 and OmniHuman (wav2vec), LatentSync and MuseTalk (Whisper), Avatar Forcing (wav2vec 2.0). The speech model already "knows" phonemes, so the video model does not need to learn them from scratch.

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

5 frames

Crop the mouth: 0.2 s of video (25 fps), plus the matching 0.2 s of audio

→

Embed

SyncNet (Oxford, 2016) maps both to vectors; in sync = close together

→

Slide

Shift the audio ±15 frames and find where the distance is smallest

→

Score

Sync-D = that smallest distance (lower is better). Sync-C = how much it stands out from the other shifts (higher is better)

Sync-C ≈ 7–8 with Sync-D ≈ 6.5–7 is close to real video on lip-reading datasets (Wav2Lip, ACM MM 2020).

Reading the scoreboard (and its trap)

System	Benchmark	Sync-C ↑	Sync-D ↓
Real video	LRS2 (lip-reading set)	7.84	6.74
Wav2Lip (2020)	LRS2	7.79	6.39 ← beats real video?
NAVA (Baidu, 2026)	Verse-Bench	7.79	7.57
Ovi 1.1 (Character.AI)	Verse-Bench	7.48	7.98
LTX 2.3 (Lightricks)	Verse-Bench	7.25	7.69
daVinci-MagiHuman	Verse-Bench	7.15	7.82

Only compare numbers on the same benchmark. Wav2Lip was trained against a SyncNet-style judge, so it can “beat” reality on that judge. Sources: Wav2Lip arXiv 2008.10010; NAVA arXiv 2605.30073.

Game: which one is in sync?



Clip A

Watch his mouth on the hard consonants.

Clip A (re-timed from Veo 3) · click to watch · source ↗



Clip B

Same clip. Is anything different?

Clip B (re-timed from Veo 3) · click to watch · source ↗



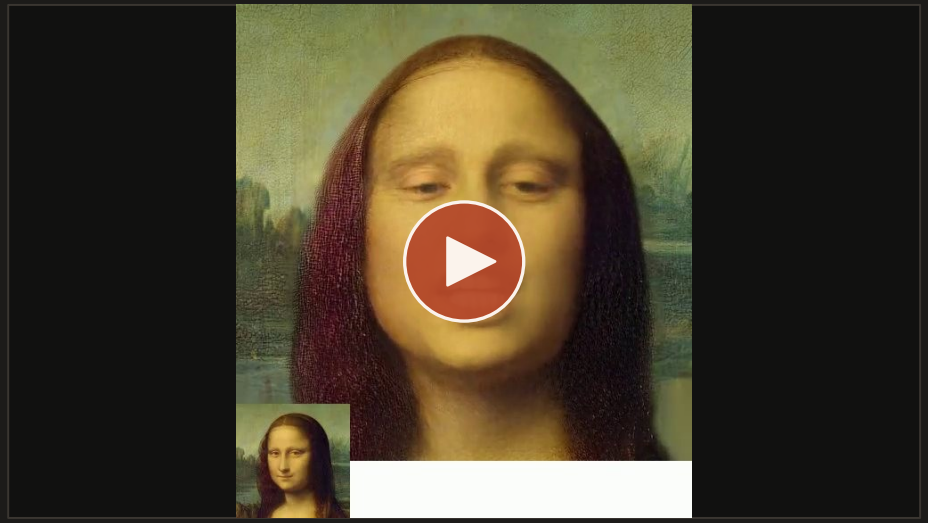
Clip C

And this one?

Clip C (re-timed from Veo 3) · click to watch · source ↗

Vote: A, B or C? Then: is any of them early rather than late?

VASA-1: a painting that talks



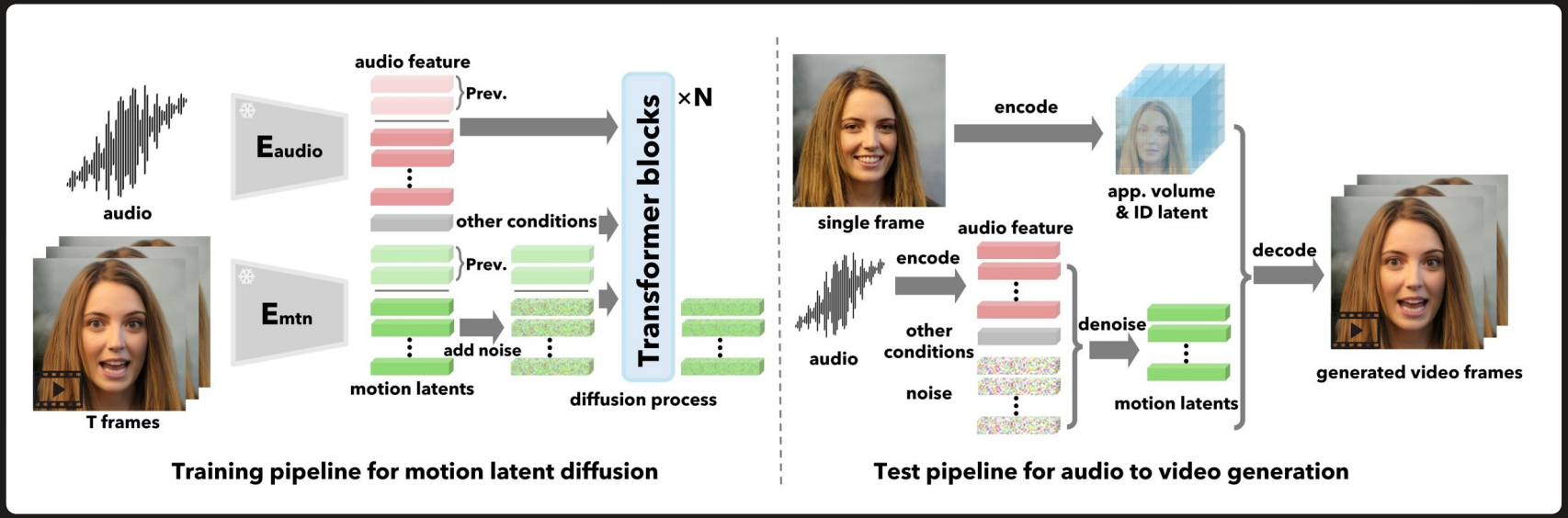
VASA-1 · Microsoft Research · click to watch

- One photo + audio → a talking face
- Generates motion, not pixels (~29M parameters)
- 40 fps, ~170 ms on one RTX 4090
- Not released: misuse risk

▶ VASA-1 project page

vasavatar.github.io/VASA-1

How VASA-1 works: generate motion, not pixels



Left: training a diffusion transformer on motion latents, conditioned on audio and other signals. Right: one photo is encoded once into appearance and identity; only the motion is generated, then decoded into frames. Figure: VASA-1, Microsoft Research, arXiv 2404.10667.

Same face, same words, four emotions



VASA-1 lets you dial gaze, head distance and emotion as separate controls · VASA-1 · Microsoft Research · click to watch · source ↗

OmniHuman: from face to whole body



OmniHuman-1 · ByteDance · click to watch

- One image + audio → full body
- 18.7K hours of mixed-condition data
- v1.5: an LLM plans the acting

▶ **OmniHuman-1 project page**

omnihuman-lab.github.io

▶ **OmniHuman-1.5 (arXiv 2508.19209)**

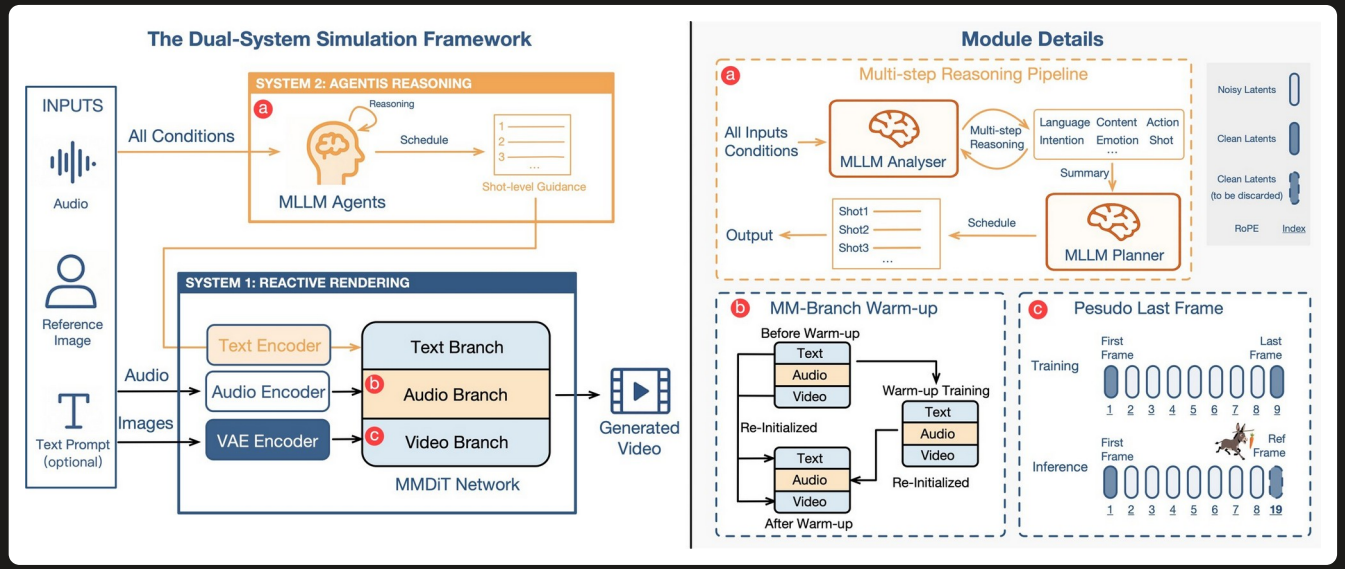
arxiv.org/abs/2508.19209

OmniHuman-1.5: acting, not just talking



Watch the eyes and the pauses, not only the lips · OmniHuman-1.5 · ByteDance · click to watch · source ↗

OmniHuman-1.5: a planner thinks, a renderer acts



"System 2": multimodal LLM agents analyse the audio, image and text and write a shot-by-shot plan. "System 1": an MMDiT with text, audio and video branches renders it. Figure: OmniHuman-1.5, ByteDance, arXiv 2508.19209.

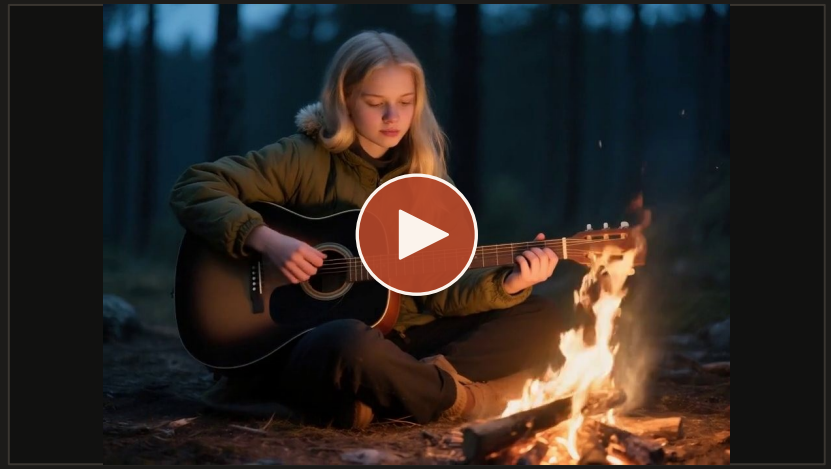
Singing: voice drives breath, body and instrument



Wan2.2-S2V (Alibaba)

Speech-to-video for film: a singer on a ship in a storm, camera and body moving with the song.

Wan2.2-S2V · Alibaba Tongyi · click to watch · source ↗

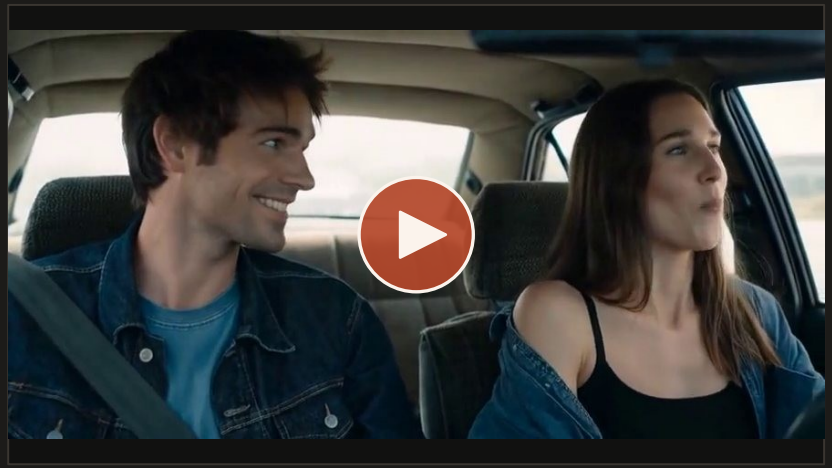


HunyuanVideo-Avatar (Tencent)

Strumming hands, breathing and firelight from one photo and a song.

HunyuanVideo-Avatar · Tencent · click to watch · source ↗

Two people: who is speaking?



MultiTalk (NeurIPS 2025)

Each voice track is bound to one person, so the right mouth moves.

MultiTalk · MeiGen-AI · click to watch · source ↗

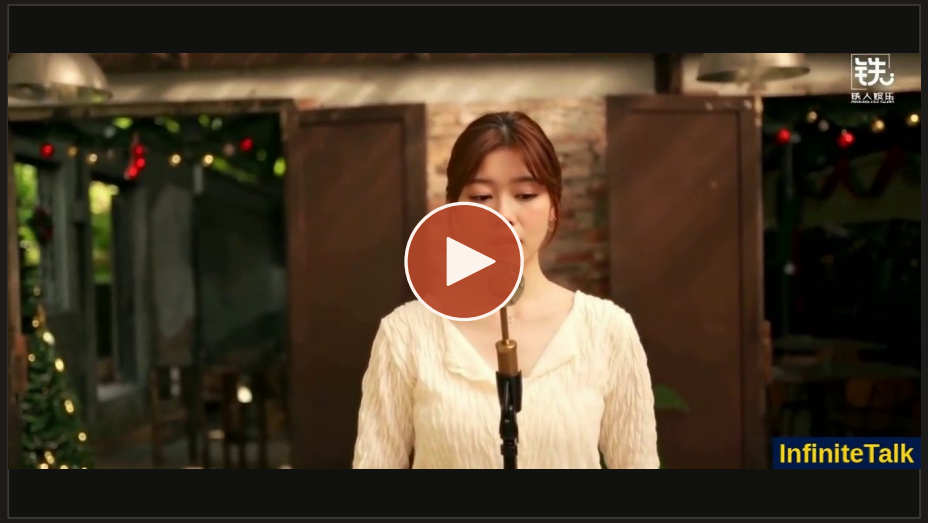


HunyuanVideo-Avatar (Tencent)

Multiple characters, with an emotion module for each voice.

HunyuanVideo-Avatar · Tencent · click to watch · source ↗

Dubbing: one performance, many languages



InfiniteTalk · MeiGen-AI · click to watch

- Re-generates lips, head and body to a new voice
- Open weights; any length
- Film precedent: Fall (2022), 30+ words “vubbed”

▶ InfiniteTalk project page

meigen-ai.github.io/InfiniteTalk

▶ Flawless TrueSync on Fall (YouTube)

www.youtube.com/watch?v=iQ1OPpj8gPA

A performance built from references



Character, wardrobe, set and voice can each be pinned by a reference · Vidu · ShengShu (version not stated) · click to watch · source ↗

Directing with references



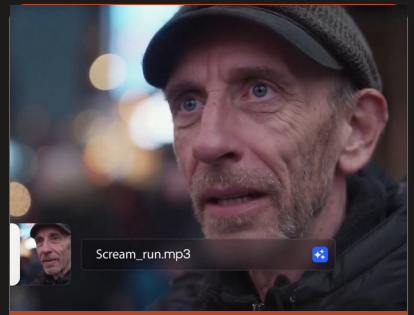
Who

Identity refs: up to 15 images (Vidu Q4)



How they sound

Voice refs: Vidu Q4, Seedance 2.0



Your own track

Record the line; LTX-2 or Wan-S2V acts it



Whose line

One track per person (MultiTalk, NAVA)

The more you pin by reference, the less you leave to chance, and the more consent you need.

Directing the delivery: what to ask for

Direction	Voice	Face and body	Prompt words that help
Tired	Slower, lower, breathy	Heavy lids, slow blinks, little head motion	“quietly, exhausted, after a long pause”
Angry, held in	Clipped, even pitch	Tight jaw, fixed stare	“through clenched teeth, without raising his voice”
Nervous	Fast, rising ends, fillers	Gaze darts away, small swallows	“hesitant, um, avoids eye contact”
Delighted	Wider pitch range, laugh in voice	Raised cheeks, eyes crinkle	“can’t hide a smile, laughs on the last word”

Actors call these “playable actions”. Describe what the character does, not only the emotion label.

EXERCISE · 5 MINUTES

One line, three deliveries: “I never said you could leave.”

In pairs, write three prompts that make this line a threat, a joke, and a plea. Describe the voice, the face and one gesture for each.

Performers have been digitised before

2002 — Andy Serkis performs Gollum through motion capture

2021 — Roadrunner synthesises three lines in Anthony Bourdain's voice; uproar

2022 — James Earl Jones licenses his archived voice for Darth Vader (Respeecher)

2022 — ABBA Voyage: ILM avatars from ~160 cameras; 4M+ visitors by 2026

2025 — The Brutalist: AI-refined Hungarian vowels spark debate

Break: student presentations

Presentations

This week's sign-ups, ~10 minutes each

10-minute break

Then Story 3: characters that listen

Try it

Any demo, sound off then on: what changed?

STORY 3 OF 4

03

Characters that listen

From rendering a clip to holding a conversation, in real time



Two numbers that define “real time”

25 fps

Throughput: make video at least as fast as it plays. 40 ms per frame

~200 ms

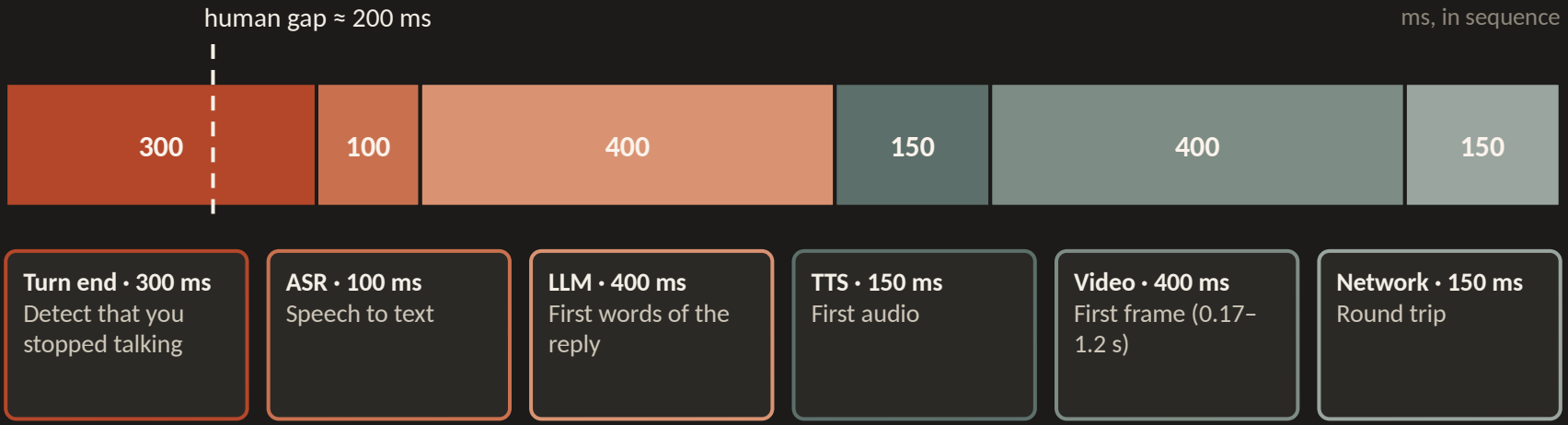
Latency: the typical gap between turns in human conversation

600 ms+

What it takes a human to plan a spoken reply, so we predict the end of your sentence

A model can be fast (high fps) and still feel slow (high latency). Conversation needs both. Source: Levinson & Torreira, Frontiers in Psychology 2015; Stivers et al., PNAS 2009 (10 languages).

The latency budget of a talking avatar



Run in sequence, a typical pipeline takes about 1.5 s, far from 200 ms. Fixes: overlap the stages (stream every step), predict the end of the turn, and make the video model answer in chunks. Voice-agent figures: Soniox latency guide; video range from the papers on the next slides.

Why a normal video diffusion model can’t talk back

50 steps

A typical sampler denoises 30–50 times

→

× 2

Classifier-free guidance runs each step twice (Week 3)

→

× all frames

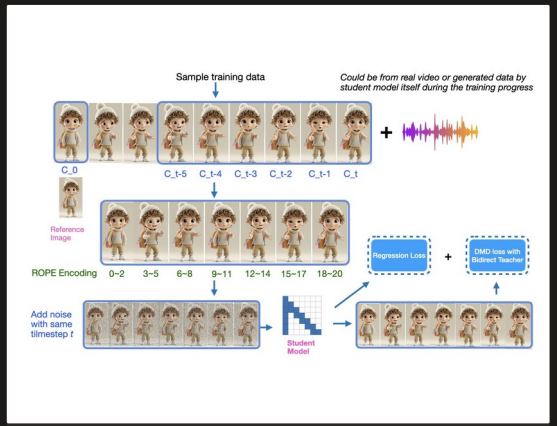
Every frame attends to every other, past and future, so nothing is ready until the whole clip is

→

Seconds to minutes

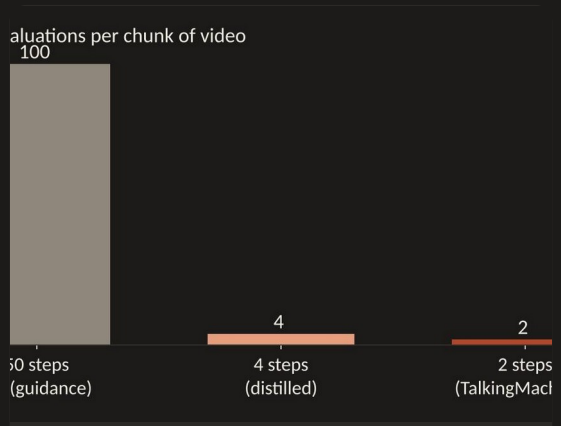
Fine for a film; useless for a conversation

Three tricks you already know



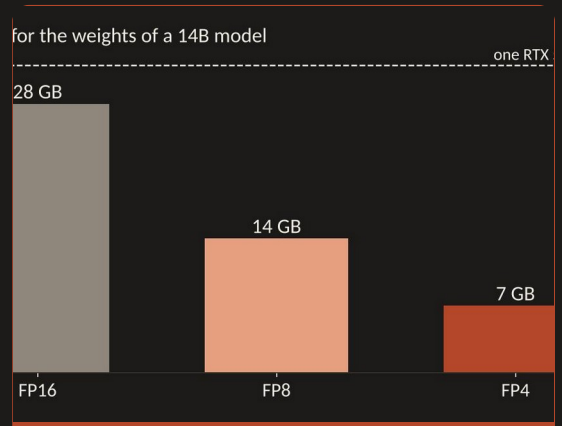
Chunks + KV cache (Week 7)

A few frames at a time; cache the past



Fewer steps

100 model passes → 2 by distillation

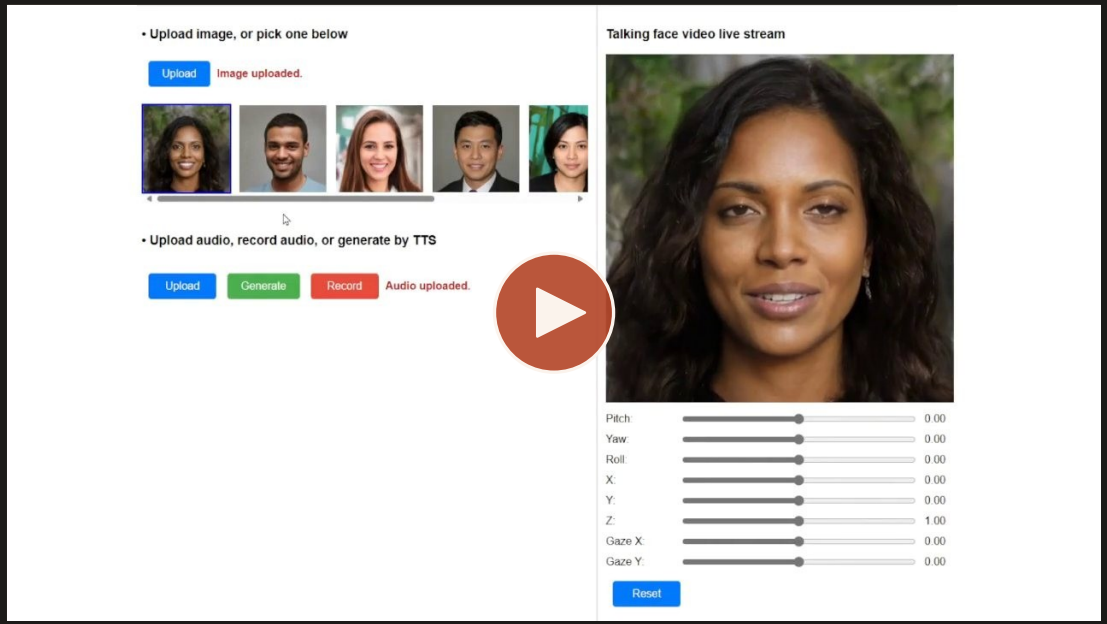


Fewer bits (Week 6)

14B weights: 28 GB → 7 GB at FP4

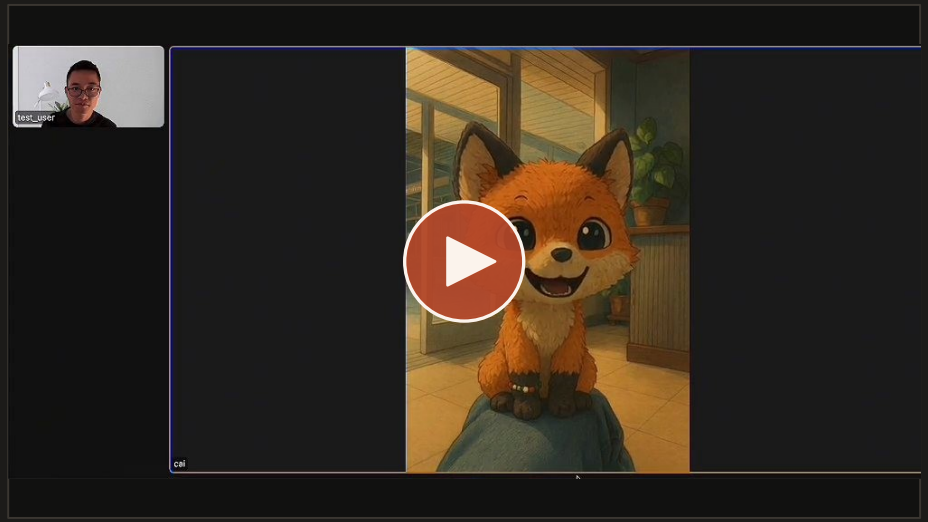
Plus a systems trick: run the transformer and the decoder on different GPUs in parallel.

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



Up to 40 fps at 512×512 with ~170 ms delay, on one RTX 4090 · VASA-1 · Microsoft Research · click to watch · source ↗

TalkingMachines: a FaceTime call with a character



TalkingMachines · Character.AI · click to watch

- Wan 2.1 → an 18B real-time avatar
- 24 → 2 model passes per chunk (DMD)
- Cache keeps the first chunk (identity) + the last

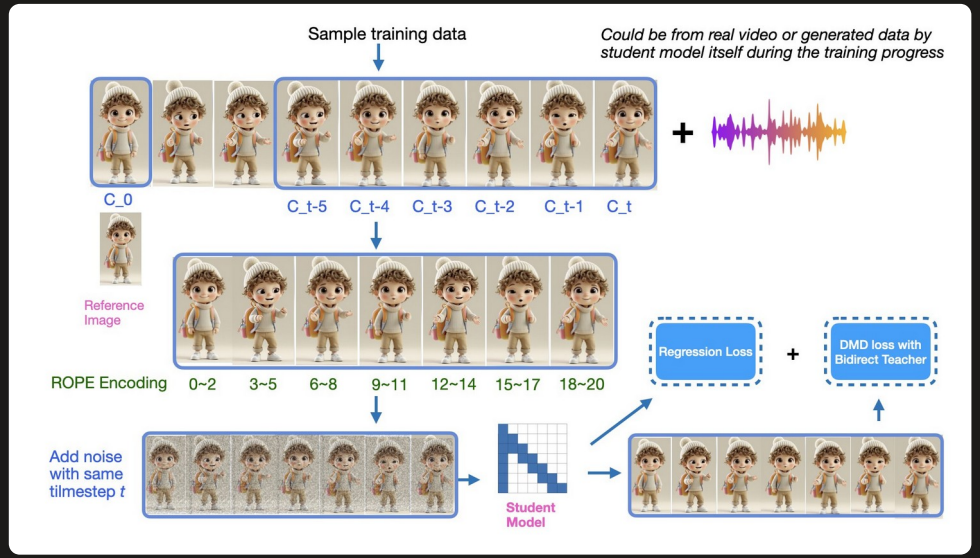
▶ **TalkingMachines project page**

aaxwaz.github.io/TalkingMachines

▶ **Paper (arXiv 2506.03099)**

arxiv.org/abs/2506.03099

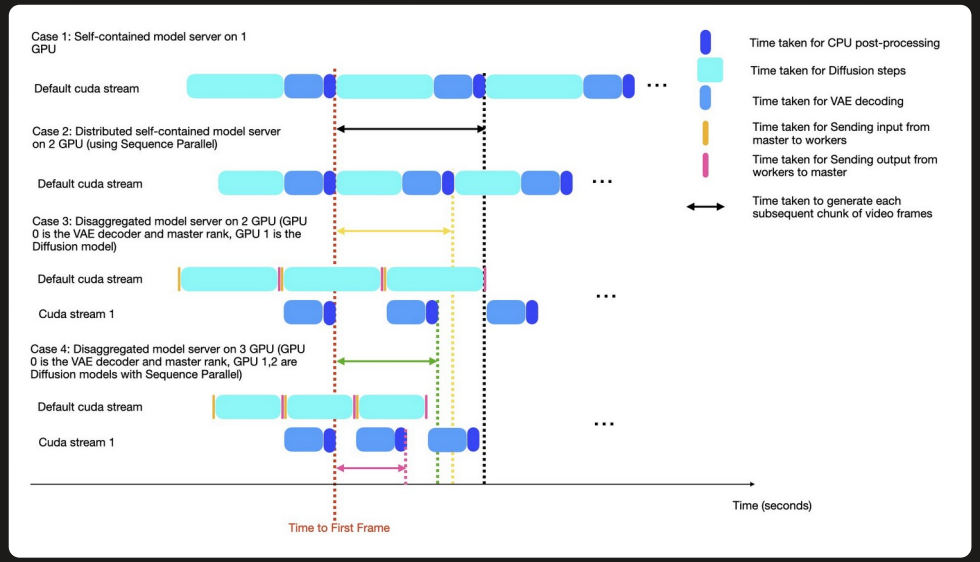
How streaming distillation works (TalkingMachines)



- Top: training clips are cut into chunks of 3 latent frames, plus the reference image (C_0)
- Each chunk only sees itself, the first chunk and the previous chunk: a sparse causal mask
- A slow bidirectional teacher (24 steps) supervises a fast causal student (2 steps)
- Training can even use the student's own generated video, like Self Forcing (Week 7)

Figure: TalkingMachines, Character.AI, arXiv 2506.03099 (2025).

Where the milliseconds go: split the work across GPUs



- Case 1: one GPU does diffusion, then VAE decoding, then waits
- Case 3: GPU 0 decodes (VAE) while GPU 1 already denoises the next chunk
- Case 4: more GPUs share the diffusion steps (sequence parallel)
- The red line, time to first frame, moves earlier with each design

Figure: TalkingMachines runtime analysis, Character.AI, arXiv 2506.03099.

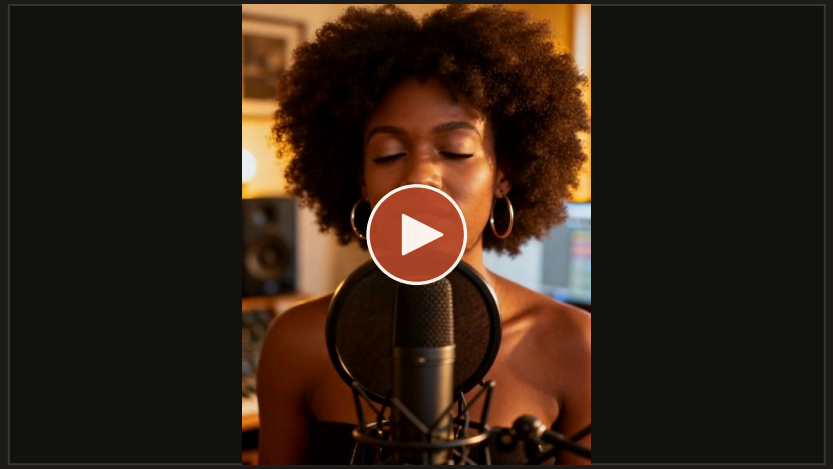
StreamAvatar: the whole upper body, live



Talking

Distilled from Wan2.2 into a 3-step streaming model: 25 fps, ~1.2 s delay on two H800s.

StreamAvatar · Tencent Hunyuan et al. · click to watch · source ↗

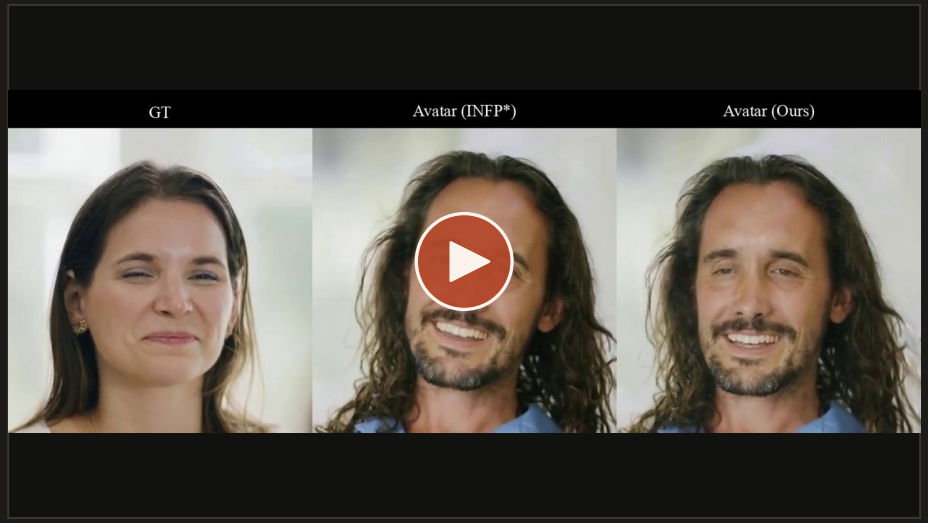


Singing

A “reference sink” keeps the first frames in memory so identity does not drift.

StreamAvatar · Tencent Hunyuan et al. · click to watch · source ↗

Avatar Forcing: an avatar that listens (CVPR 2026)



Avatar Forcing · KAIST, NTU · click to watch

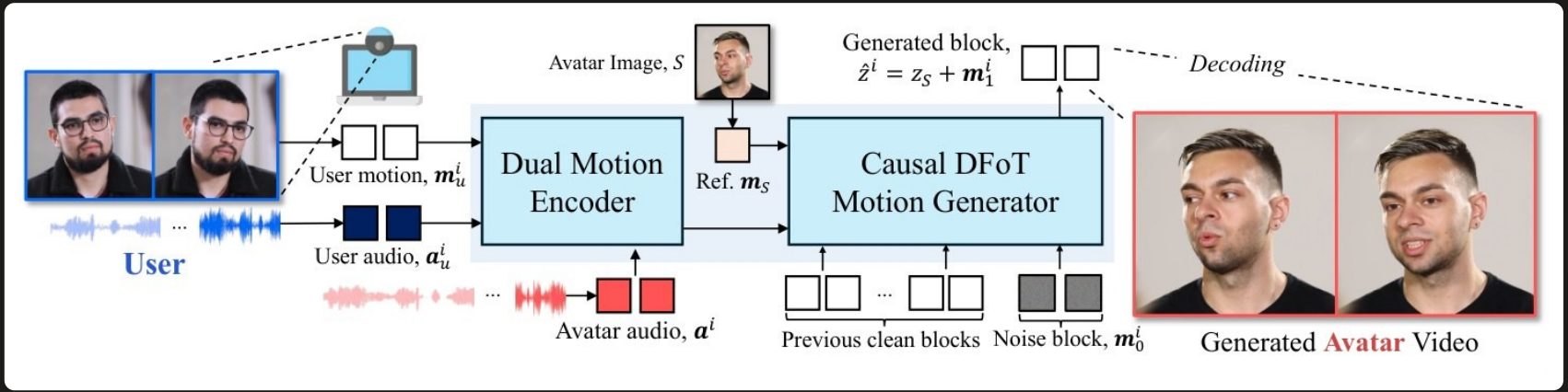
Columns: real person · previous method · Avatar Forcing

- Reacts to your voice and motion while you talk
- 10-frame blocks, rolling KV cache
- ~500 ms on one H100; preferred >80%

▶ **Project page (more listening demos)**

taekyungki.github.io/AvatarForcing

Inside Avatar Forcing



The user's motion and audio and the avatar's own audio are encoded together (Dual Motion Encoder); a causal diffusion-forcing generator writes the next block of motion, conditioned on past clean blocks and a reference frame; a decoder turns motion into video. Figure: Avatar Forcing, arXiv 2601.00664 (CVPR 2026).

Real-time scoreboard, 2024–2026

System	Delay to respond	Speed	Hardware	Open?
VASA-1 (Microsoft, 2024)	~170 ms	40 fps	1× RTX 4090	No
Livatar-1 (Hedra, 2025)	0.17 s	141 fps	1× A10	No
Avatar Forcing (CVPR 2026)	~0.5 s	25 fps	1× H100	Page only
OmniForcing (2026), with sound	~0.7 s	~25 fps	1 GPU	Paper
Hallo-Live (2026), with sound	0.94 s	20 fps	2× H200	Yes
StreamAvatar (2025), body	~1.2 s	25 fps	2× H800	Not stated
Wan-Streamer v0.1 (2026)	~0.2 s + network	25 fps (192p)	2 GPUs	Not stated

Not directly comparable: resolutions, face-only vs body, and what counts as “delay” differ by paper. Sources: each paper’s arXiv page (see reading guide).

Pipeline or one brain?

Cascade (most products today)

- Separate models: turn detection → ASR → LLM → TTS → avatar
- Easy to swap parts and pick the best of each
- Every hand-off adds delay and loses tone of voice
- Typical: around 1 s or more

End-to-end (Wan-Streamer, 2026)

- One transformer takes in text, audio, video and puts out audio + video
- Streams in 160 ms units; a “thinker” passes its memory to a “performer”
- Can hear how you said it, not only what you said
- ~200 ms model-side; still low resolution

Wan-Streamer: one model hears, thinks and answers



A live call: the person in the small window talks; the character answers in voice and video · Wan-Streamer · Alibaba Wan · click to watch · source ↗

From our lab: TalkVid

1,244 h

of talking-head video, filtered from 6,000+ hours of web video

7,729

speakers, ages 0 to 60+, upper body included

15

languages (867 h English, 249 h Chinese); 87.6% at 1080p or above

500

TalkVid-Bench clips, balanced by age, gender, ethnicity and language

TalkVid: A Large-Scale Diversified Dataset for Audio-Driven Talking Head Synthesis (CVPR 2026 Findings). CUHK-Shenzhen, SYSU and HKUST. arXiv 2508.13618 · github.com/FreedomIntelligence/TalkVid

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

Collect

30,000+ videos at 1080p or higher

→

Cut

Shot detection; drop clips under 5 s or without speech

→

Look

Aesthetic score (DOVER ≥ 7) for sharp, well-lit video

→

Move

Motion tracking: drop jittery and frozen clips

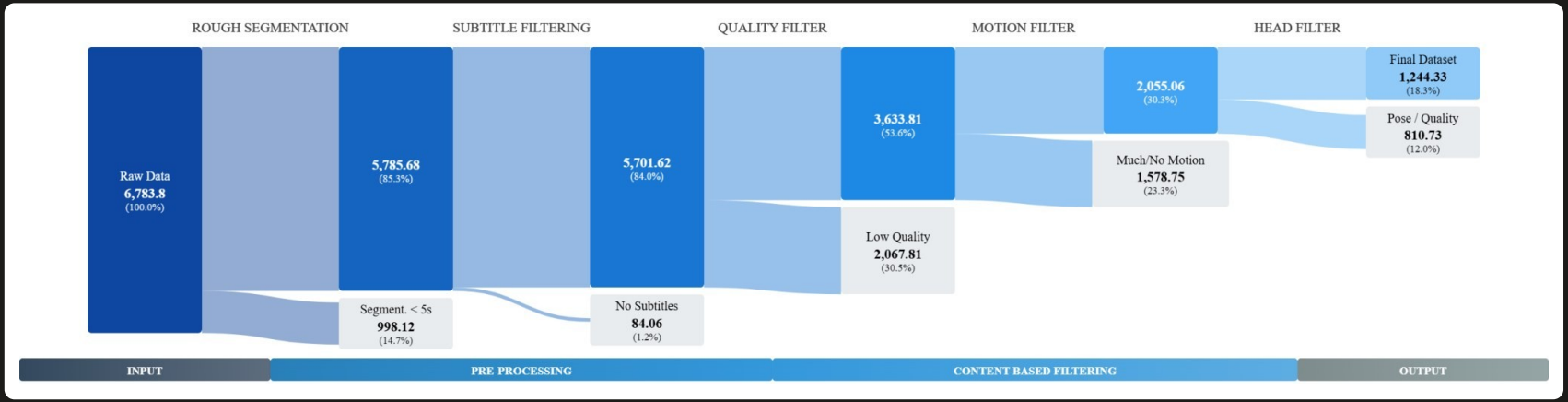
→

Check

Head pose and completeness; humans agree with the filters 95% of the time

Training on the balanced 160-hour TalkVid-Core gave lower FVD than training on older datasets (175.1 vs 188.7 HDTF / 183.2 Hallo3) on the HDTF test set.

The real filter funnel, and what it keeps



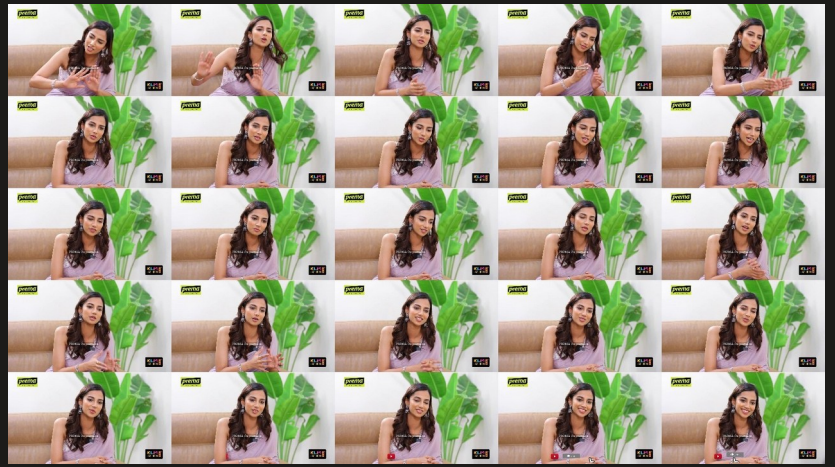
Raw hours on the left; segmentation, subtitle, quality, motion and head filters remove clips until 1,244 hours remain. Figure: TalkVid, arXiv 2508.13618 (CVPR 2026 Findings).

What the quality filter drops and keeps



Dropped: low aesthetic score

Dim, soft, low-contrast footage.



Kept: high aesthetic score

Sharp, well lit, stable framing.

Example frames from TalkVid's DOVER filter (arXiv 2508.13618). A model learns the look of its data, so this filter decides the look of every avatar trained on it.

What runs where (for your projects)

A laptop or mid GPU

MuseTalk: single-step mouth inpainting, 30+ fps on a V100-class GPU. LatentSync 1.5: about 8 GB of GPU memory (offline).

One RTX 4090 / 5090

Light real-time heads (e.g. SoulX-FlashHead Lite, reported 96 fps on a 4090). 14B avatar models at 480p offline, not live.

Several data-centre GPUs

Live 14B-class avatars with body and sound: StreamAvatar, Hallo-Live, FlashTalk (8× H800).

Real-time 14B-class avatars on one consumer GPU is an open problem, and one our lab works on: it needs distillation, chunked generation and low-bit kernels together.

Where live characters already are

Taxis in Japan

ESTsoft's live avatar, with NTT and Nihon Kotsu (Feb 2026)

Creator tools

ElevenLabs avatars: script → talking video (Jun 2026)

Characters

Character.AI: video calls with fictional characters

Performance

ABBA Voyage: digital performers, 4M+ visitors

Artists working with their own likeness

xhairymutantx · Whitney 2024

Herndon & Dryhurst: every prompt returns a distorted “her”, a likeness the artist controls.

The Call · Serpentine 2024–25

15 UK choirs, singing models, and a data trust so singers share control.

A pattern for your own projects: use your own voice and face, or people who agreed, and make the consent part of the work.

FOR THE ROOM

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

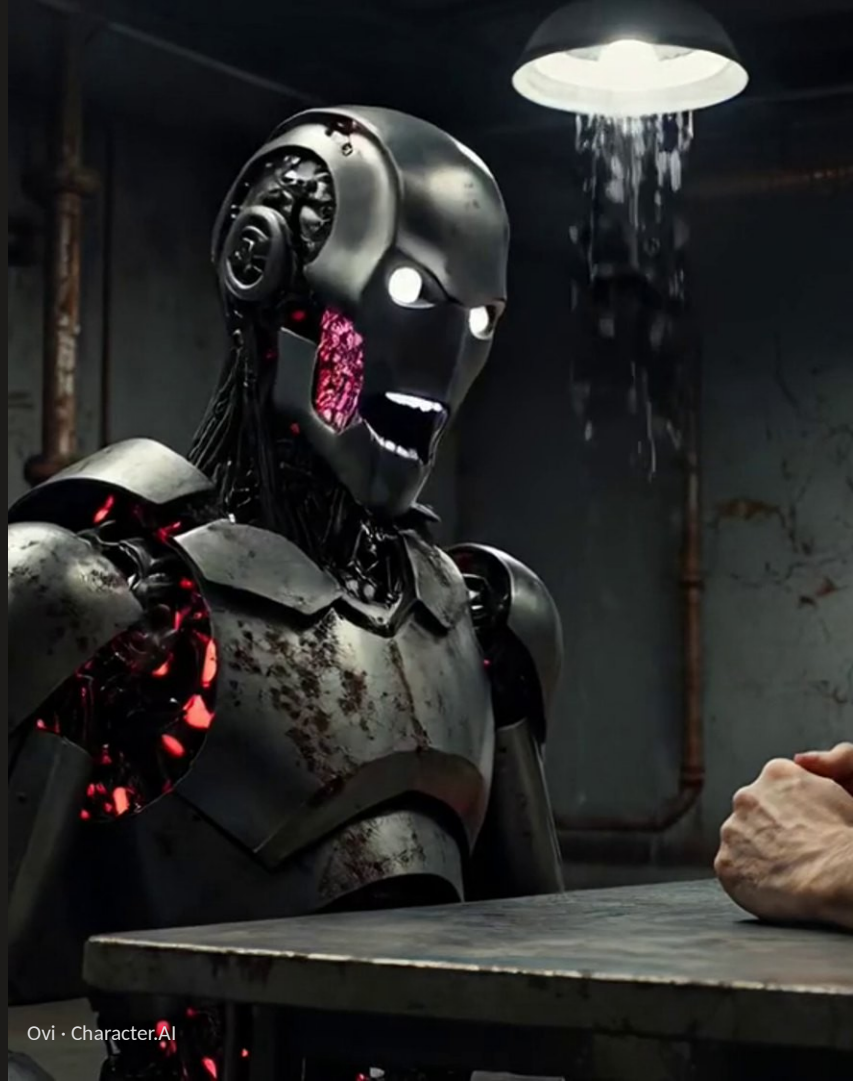
What would make it feel respectful, and what would make it feel wrong?

STORY 4 OF 4

04

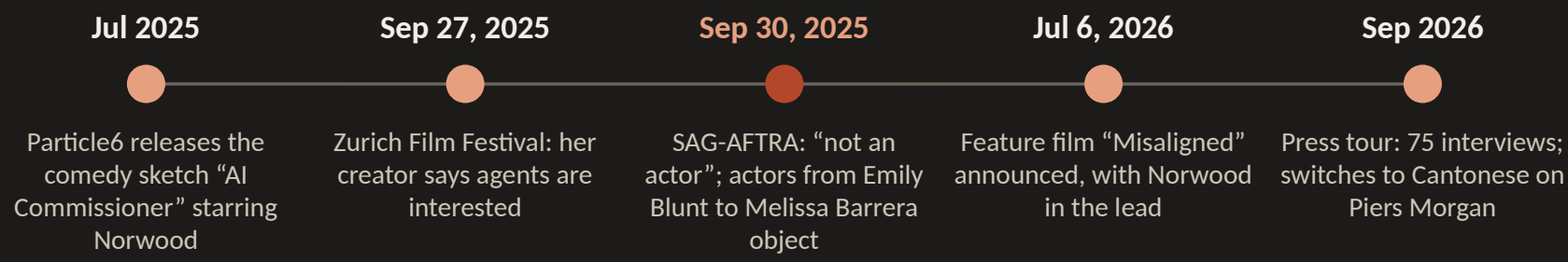
Who is the performer?

Synthetic actors, digital replicas, and the new rules of 2026



Ovi · Character.AI

Tilly Norwood: an “actress” who never existed



Watch: synthetic performers, from sketch to stage

Tilly Norwood · “AI Commissioner”

The July 2025 comedy sketch that introduced her (Particle6).

▶ Play on YouTube

Tilly Norwood · “Take The Lead”

Her March 2026 music video answering the backlash; 18 human credits, AI vocals.

▶ Play on YouTube

ABBA Voyage

Digital performers with the band’s consent and a live band: 4M+ visitors.

▶ Official trailer

Holly Herndon · Holly+

The artist’s own voice model, offered to others on her terms.

▶ “Jolene (feat. Holly+)”

Opens in the browser. Same technology, four different answers to “who agreed?”

Two positions, in their own terms

SAG-AFTRA (Sep 30, 2025)

- Norwood is a character generated by a program, not an actor
- Trained on countless performers' work without permission or pay
- Has no life experience or emotion to draw on
- Producers must notify and bargain with the union before using a synthetic performer

Her creator (Particle6)

- A creative work, "a piece of art", not a replacement for a human
- AI is a new tool, like animation, puppetry or CGI
- Says she has no intention of replacing human actors
- Plans dozens more characters in a "Tillyverse"

What the actors won: the 2026 SAG-AFTRA contract

Synthetic performers

Only with “significant additional value”; tell the union first

Digital replicas

Informed consent and fair pay, living or dead

Strikes and minors

No replica during a strike; extra protection for minors

Ratified June 4, 2026 (91.4% yes), to June 2030. “A strong disincentive, not a ban.”

Source: LA Magazine; TheWrap; Malay Mail/AFP (May–June 2026)

The law, region by region (October 2026)

Where	Rule	Status
California	AB 2602 voids vague digital-replica clauses; AB 1836 needs an estate's consent for replicas of the dead	In force since Jan 2025
United States	NO FAKES Act: a federal right to your voice and likeness, with takedowns	Not law; blocked Sep 30, 2026
Mainland China	AI-generated content must carry visible labels and hidden metadata labels	In force since Sep 1, 2025
European Union	AI Act Art. 50: deepfakes must be disclosed; outputs machine-marked	Applies from Aug 2, 2026
Denmark	Copyright-style protection of your own face and voice against deepfakes	Bill in parliament, Oct 2026

Not legal advice. For your work: label AI-generated people and voices, and get consent from anyone recognisable.

Your voice is part of your identity

Lehrman & Sage v. Lovo (2025)

“Research” recordings became clone voices; the court let their claims proceed

“Sky” (2024)

Johansson said no; a similar voice shipped; OpenAI paused it

Vader in Fortnite (2025)

The estate agreed; the union still objected

What audiences say (NRG × TheWrap, March 2026)

56%

say AI actors will never be as good as human actors

7%

say AI actors are already as good

~3 in 5

say releasing a film with an AI-written script is unacceptable

Yet the largest group said AI use would not change their interest in a film. Audiences judge the result, and the story about how it was made.

Two flashpoints from this year

Feb 2026 · Seedance 2.0 clip

Two stars' likenesses, no consent: condemned by the MPA and SAG-AFTRA

Jun 2026 · Dreams of Violets

Fully AI-generated feature premieres at Tribeca

Same tools, opposite receptions: whose likeness, whose story, who agreed.

A checklist for your Assignments 2 and 3

Consent

Real faces or voices only with written consent

Label

Mark what is generated (required in China, EU)

Credit

Models, voice refs, every human contributor

Keep sources

Prompts, seeds, audio refs, versions, raw takes

DEBATE · 10 MINUTES

“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 speakers per side, one minute each

Then the whole room votes by show of hands

Arguments to start from

For

- Animated performances already move us; awards could judge the result, not the method
- Directors, writers and voice references made the choices; the award could go to them
- A separate category could honour new craft without displacing actors

Against

- Acting awards honour a person’s lived skill and risk on screen
- The “performance” is averaged from thousands of uncredited actors
- Prizes for synthetics reward replacing the people the model learned from

Four stories, one line each

01

Sound

Picture and sound now share one model and one clock; the human job is choosing what to hear.

02

Performance

Direct with references and playable actions; check the p, b and m sounds.

03

Real time

Chunks, fewer steps and fewer bits turn a renderer into a conversation partner that listens.

04

Performer

Consent, credit and labels are now contract and law, not just good manners.

Tonight → your assignments (both due Nov 30)

Assignment 2 · Reverse, Rebuild, Remix

When you reverse-engineer an AI video, reverse-engineer its sound too: which layers were generated, and in which model? Then remix with a native-audio model or your own voice track.

Assignment 3 · Make Your Own Game

Give one character a voice and a face that reacts. A talking portrait (MuseTalk, LatentSync) or a short generated cut-scene with dialogue is enough.

Use your own voice, or a friend who agreed in writing. Label what is generated.

Three readings for this week

01

LTX-2: Efficient Joint Audio-Visual Foundation Model

Technique · Lightricks, Jan 2026. Abstract, architecture figure, two demos.

arXiv 2601.03233

02

The Call · Holly Herndon & Mat Dryhurst

Art + tech · Serpentine, 2024–25. Choirs, singing models, and a data trust.

serpentinegalleries.org

03

Tilly Norwood to lead her first film

Culture · Variety, July 2026, with the SAG-AFTRA statement.

variety.com

+

Optional: Avatar Forcing; TalkVid

Technique · real-time listening avatars, and the data behind them.

arXiv 2601.00664 · 2508.13618

Quiz 08 · five short questions

Two technique questions

Joint audio-video vs adding sound afterwards; the tricks that make avatars real-time.

Two art questions

Synchresis and why lip-sync errors are so visible; directing a line beyond the words.

One reading reflection

Any one of the three readings: its main idea, and whether you agree.

hyang.org/amcc5160/quiz-08 · 4 points each · due at the end of the semester, editable until then.

Every demo from tonight, in one place

Story	Demo	Where to find it
1 · Sound	Veo 3, Ovi, LTX-2, UniVerse-1	deepmind.google/models/veo · aaxwaz.github.io/Ovi · github.com/Lightricks/LTX-2
2 · Performance	VASA-1, OmniHuman, MultiTalk, HunyuanVideo-Avatar, InfiniteTalk	microsoft.com/…/vasa-1 · omnihuman-lab.github.io · github.com/MeiGen-AI
3 · Real time	VASA-1 live, TalkingMachines, Avatar Forcing	arxiv.org/abs/2506.03099 · taekyungki.github.io/AvatarForcing
4 · Performer	Tilly Norwood, SAG-AFTRA, The Call	variety.com · sagaftra.org · serpentinegalleries.org

All links are also on the Lecture 08 page, under “Demos”. Videos belong to their creators and are shown for teaching.

Next week: critique, evaluation, provenance

Metrics vs eyes

What do scores really measure?

Studio critique

Bring one generated clip with sound

Provenance

Model cards, watermarks, labels

Sound, performance, presence

Generated video learned to speak, to act, and to listen. Choose what we hear, direct how it is said, and ask whose voice it is.

Sources: Google DeepMind Veo 3 model card; LTX-2 (arXiv 2601.03233); Ovi (arXiv 2510.01284); Kandinsky 6.0 (arXiv 2610.05608); NAVA (arXiv 2605.30073); OmniForcing (arXiv 2603.11647); Wav2Lip (arXiv 2008.10010); VASA-1 (arXiv 2404.10667); OmniHuman (arXiv 2502.01061, 2508.19209); TalkingMachines (arXiv 2506.03099); Avatar Forcing (arXiv 2601.00664); Livatar-1 (arXiv 2507.18649); Wan-Streamer (arXiv 2606.25041); TalkVid (arXiv 2508.13618); CMU 10-423 Lecture 24; Levinson & Torreira 2015; SAG-AFTRA, Variety, TheWrap, LA Magazine, NRG. Demo videos belong to their creators, shown for teaching.

APPENDIX

A

Do it yourself

Tools, a workflow, prompts and checks for Assignments 2 and 3

Toolkit: open models for each job

Job	Open model	Size / hardware (as reported)	Licence
Picture + sound together	LTX-2 (Lightricks)	19B; one large GPU; ComfyUI support	LTX-2 Community
Picture + sound, lighter	Kandinsky 6.0 Lite · NAVA	3B · 6.3B	MIT · Apache-2.0
Two-person dialogue	Ovi · MultiTalk	11B · 480p on one RTX 4090 (low-VRAM mode)	Apache-2.0
Add Foley to a silent clip	MMAudio · HunyuanVideo-Foley	MMAudio from 157M: 8 s of audio in ~1.2 s	MMAudio weights non-commercial
Lip-sync an existing video	LatentSync 1.5 · MuseTalk	~8 GB GPU memory · 30+ fps (256×256 face)	Apache-2.0 · open
Re-dub a whole video	InfiniteTalk	480p/720p; unlimited length	Apache-2.0
Talking or singing body	Wan2.2-S2V · HunyuanVideo-Avatar	14B · 24 GB min (10 GB with TeaCache)	Apache-2.0 · Tencent licence

Check each licence before commercial use. Closed options with native sound: Veo 3.1, Seedance 2.0, Kling, Vidu, MiniMax (via their apps or APIs).

Workflow: a talking character in one evening

1 · Voice

Record the line yourself on your phone, in a quiet room. Timing is your performance

→

2 · Face

Generate or paint one image of the character, front-facing, mouth visible, good light

→

3 · Animate

Audio-to-video (LTX-2, Wan2.2-S2V, InfiniteTalk) or lip-sync (LatentSync, MuseTalk)

→

4 · Room

Add ambience and Foley (MMAudio) under the voice; keep the voice on top

→

5 · Check

Run the review checklist; fix and re-generate only the failing shot

Your own voice avoids every consent problem in Story 4, and gives you exact control of timing.

Review checklist for a generated performance



Lips

Full closure on p, b, m



Timing

Scrub to a consonant: right frame?



Voice

Same voice across shots?



Body

Gaze motivated? Hands in rhythm?



Room

Reverb fits; ambience continues



Credits

Models, refs, consent, AI label

Measure sync yourself (two commands)

Score lip-sync with the reference SyncNet (github.com/joonson/syncnet_python)

```
sh download_model.sh
python run_pipeline.py --videofile my_clip.mp4 --reference take1 --data_dir out/
python run_syncnet.py  --videofile my_clip.mp4 --reference take1 --data_dir out/
# prints AV offset (frames), min distance (Sync-D) and confidence (Sync-C)
```

Feel a delay: make a copy with the sound 125 ms late (ffmpeg)

```
ffmpeg -i my_clip.mp4 -af "adelay=125:all=1" -c:v copy late_125ms.mp4
# compare the two side by side; try 45 ms early and 125 ms late
```

If the reported offset is not 0, shift your audio by that many frames (1 frame = 40 ms at 25 fps) and re-check.