

AMCC 5160 • LECTURE 07 • 13 OCTOBER 2026

What's New in *AI Motion*

Music, minutes, and machines • Fall 2026

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

Four stories from this year

01

The 5-second clip is over

30-second single generations; a real-time minute

Mar – Sep 2026

02

Music becomes the choreographer

A song and a photo give you a minute of dance

Jul 13, 2026

03

Motion leaves the screen

A humanoid learns a K-pop dance from phone video

Jun 10, 2026

04

Sora is gone

What a platform shutdown means for artists

Apr 26 & Sep 24, 2026

Where we've been: six 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: why does it produce this?

04 • Sep 22

AI video generation and art filmmaking

05 • Sep 29

Adaptation, workflows and authorship

06 • Oct 6

Quantization for faster inference (Yaofu)

Tonight pulls these threads together: length (Weeks 3–4), speed (Week 6) and authorship (Week 5).

How many shots is this?

Watch the clip. Count the cuts. Then guess how many times the model was run.

▶ [Play: Seedance 2.5 showcase \(official page, hero video\)](https://seed.bytedance.com/en/seedance2_5)

seed.bytedance.com/en/seedance2_5

How top courses treat what's new

Course	What it does with new work	Tonight
Stanford CS348K Fatahalian · Spring 2026	A full lecture on how distillation and caching made autoregressive video real-time; guests from WorldLabs, Roblox, NVIDIA	Story 1
CMU 16-726 Jun-Yan Zhu · Spring 2025	Put a same-year fast video paper on the reading list; ends with forensics and data attribution	Story 2
MIT 4.562 Nagakura · Fall 2025	Listed Sora as the video tool in the syllabus	Story 4
USC IML-460 Holly Willis · Spring 2026	Pairs every tool with the controversy around it	Each “For the room”

STORY 1 OF 4

01

The 5-second clip is over

From fragments to scenes, and from rendering to real time

The unit of AI video is now the scene



Durations are per single generation, as stated by vendors or reviewers.

Who shipped what

Model	What's new	When
Seedance 2.5 • ByteDance	Built for 30-second storytelling; camera moves and performance blocking	Jul 31, 2026
Wan 3.0 • Alibaba	30-second clips; beta only, no open weights	Public beta, Aug 2026
Kling 4.0 • Kuaishou	Announced 3–30 s, up to 4K; Flash tier 20 s at 720p	Flash Sep 28; full model October
Helios • Peking Univ. + ByteDance	14B model; ~60 s at 19.5 FPS on one H100; open weights	Mar 2026

[Seedance 2.5 official page ↗](#)

seed.bytedance.com/en/seedance2_5

Why five seconds was the limit

6× the frames

At 24 fps, 5 s is 120 frames. 30 s is 720 frames.

~36× the attention

Full attention compares every token with every other one, so six times the tokens means roughly 36 times the work.

Errors pile up

The longer the video, the more chances for small mistakes to grow. That is drift, and we'll come back to it.

The 36× is a rule of thumb for full attention; real systems use many tricks to beat it.

Two ways to make a long video

All at once

- Denoise every frame of the clip together
- Strong consistency inside the clip
- Cost grows fast with length; you wait for the whole thing

Forward in time (autoregressive)

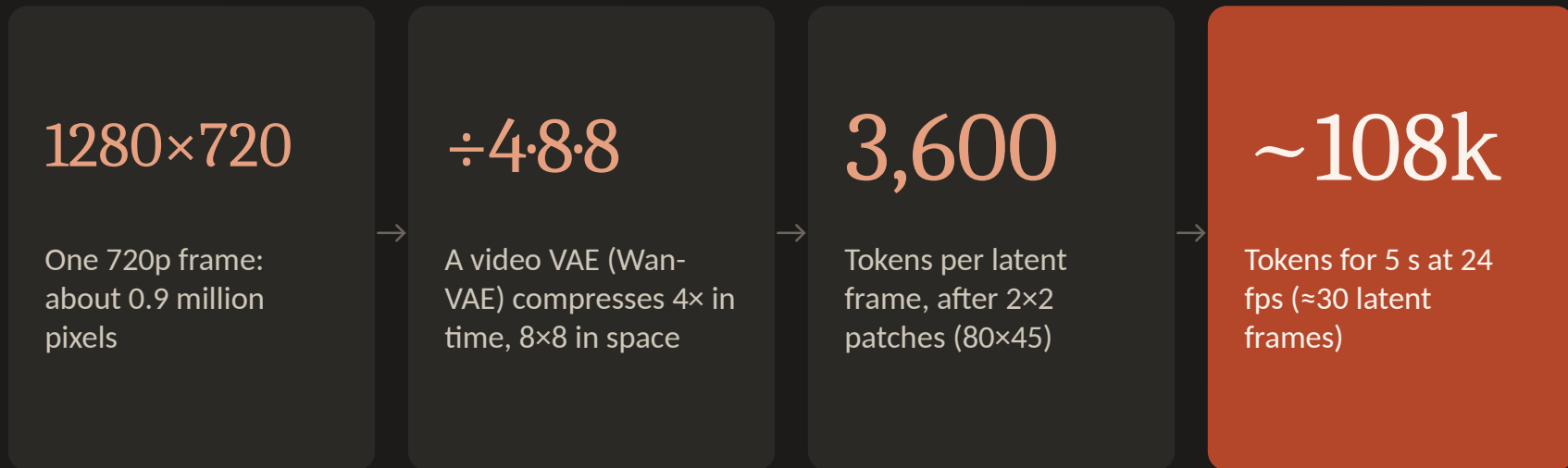
- Write the next chunk from the chunks before it
- Can stream and run in real time
- Must fight drift: errors carry forward

Helios is the second kind. Language models write text the same way: one piece after another.

Under the hood: where top courses teach this

Course	What it does with new work	Tonight
MIT 6.S184 Flow matching & diffusion · 2026	Builds a latent diffusion model from scratch; case study of large image and video generators	Tokens, flow
Stanford CS348K Fatahalian · Spring 2026	“Performance optimizations for autoregressive video models”: distillation, KV caching, token caching	Real time
CMU 16-726 Jun-Yan Zhu · Spring 2025	Video synthesis readings from Video Textures to CausVid; DMD and consistency models	Drift, steps
Stanford CS231n Fei-Fei Li et al. · Spring 2026	Opens with “generative and interactive visual intelligence”	Big picture

Under the hood: from pixels to tokens



Worked example with Wan's published settings (VAE 4×8×8, patch 1×2×2). 30 s gives ≈650k tokens.

Under the hood: why length is expensive



Rule of thumb for full attention. Real systems use windowed or sparse attention, caching and compression to beat it.

Under the hood: diffusion and flow matching

Diffusion (denoising)

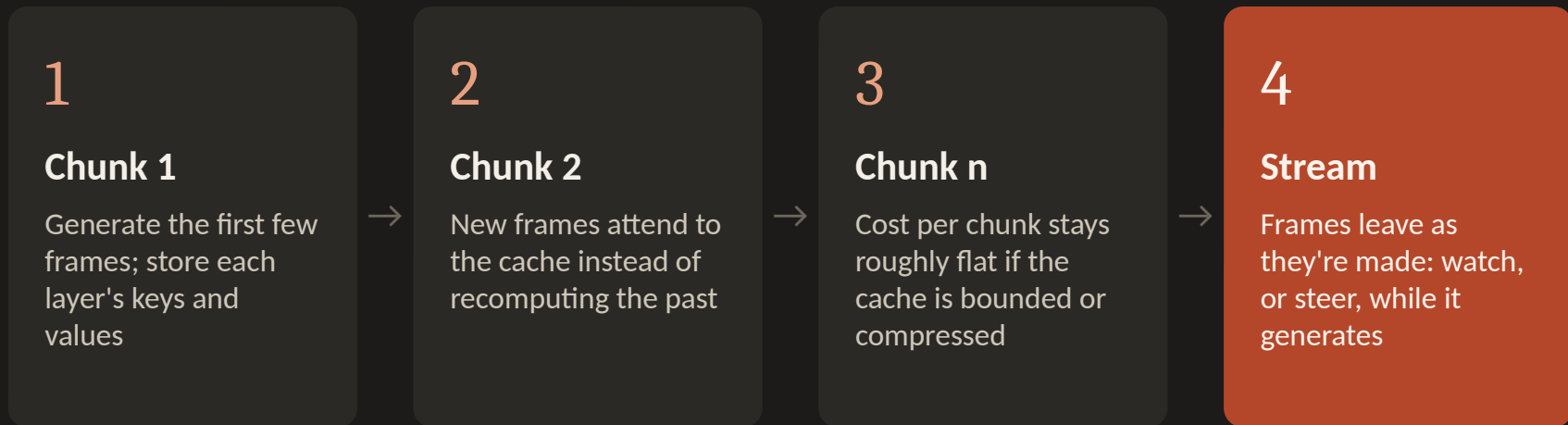
- Start from noise; remove a little noise per step
- Learns to predict the noise (or the clean video)
- Classic samplers need tens of steps

Flow matching

- Learns a velocity: which way to move from noise toward data
- Straighter paths, so fewer steps
- The backbone of many 2025–2026 video models

MIT 6.S184 teaches both from first principles. Week 3's “guidance” works the same way in both.

Under the hood: the KV cache



The same trick makes chatbots fast. Stanford CS348K (May 2026) covers it for video.

Helios: a minute of video, in real time



Support

• Text-to-Video

• image-to-Video

Frame from the official Helios feature video

- 14B autoregressive model: writes video forward, chunk by chunk
- ~60 s (1,440 frames) per generation
- 19.5 fps on one H100: close to 24 fps playback
- Open weights

▶ [Play: Helios feature video \(YouTube\)](#)

www.youtube.com/watch?v=vd_AgHtOUFQ

▶ [Minute-long demos \(project page\)](#)

pku-yuangroup.github.io/Helios-Page

[Try it: Hugging Face demo ↗](#)

huggingface.co/spaces/BestWishYsh/Helios-14B-RealTime

How Helios gets its speed

Remember less, smartly

It compresses the history it carries forward, so each new chunk doesn't re-read the whole past.

Cheaper than its size

The team reports per-step cost comparable to a 1.3B model, despite having 14B parameters.

Fewer steps

A distilled version uses 3 denoising steps instead of dozens (Week 6's theme: same model, less work).

Lean to train

Up to four 14B models fit in 80 GB of GPU memory during training, without special parallelism frameworks.

▶ [Speed demo \(project page, “Inference Speed”\)](#)

pku-yuangroup.github.io/Helios-Page

Under the hood: distillation, from 50 steps to 3

Teacher

- The original model: slow, high quality
- Tens of denoising steps per frame
- Runs offline, during training

Student

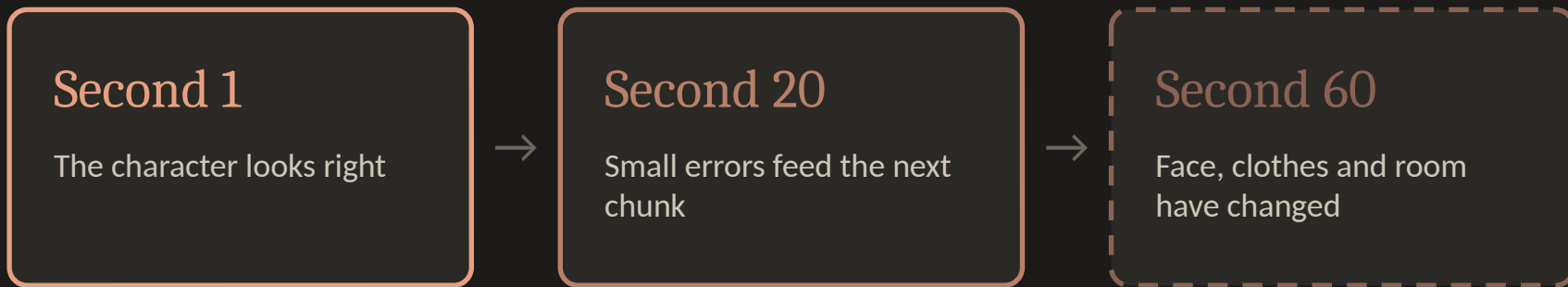
- Learns to jump where the teacher would end up
- 1–4 steps (Helios-Distilled: 3)
- Methods: consistency models, distribution matching (DMD)

CMU 16-726 lists both DMD and consistency models. Week 6's quantization cuts cost per step; distillation cuts the number of steps.

Under the hood: a real-time budget

Lever	What it cuts	Where you saw it
Deadline	24 fps leaves ~42 ms per frame; Helios's 19.5 fps is ~51 ms	Tonight
Fewer steps	Distillation: tens of steps → 3	CMU 16-726, Helios
Cheaper steps	Quantization: fewer bits per weight	Week 6 (Yaofu)
Less context	KV cache, compressed history, sparse attention	Stanford CS348K, Helios
Smaller tokens	Stronger VAE compression, bigger patches	MIT 6.S184, Wan

Why long video falls apart: drift



Helios's fix: train on deliberately drifted video, so the model learns to recover instead of patching drift at playback time.

Under the hood: the train-test gap

Teacher forcing

Train each chunk on perfect, real previous frames. At test time the model only has its own imperfect frames: exposure bias.

Diffusion Forcing · 2024

Train with different noise levels on different frames, so the model is used to imperfect context.

CausVid and Self Forcing · 2025

Distil a fast causal generator; Self Forcing (Adobe, UT Austin) trains on its own rolled-out frames with a KV cache.

Helios · 2026

Simulates drift during training instead of relying on these anti-drift heuristics at all.

Test it yourself: does she come back the same?

On the Helios project page, find “Verify Consistency”: a subject leaves the frame and returns. Then try “Interactive-Video Showcases”, where the prompt changes while the video keeps going.

▶ [Play: Helios consistency and interactive demos](#)

pku-yuangroup.github.io/Helios-Page

Real time changes what video is

From watching to playing

If frames arrive faster than you watch them, you can steer the video as it plays.

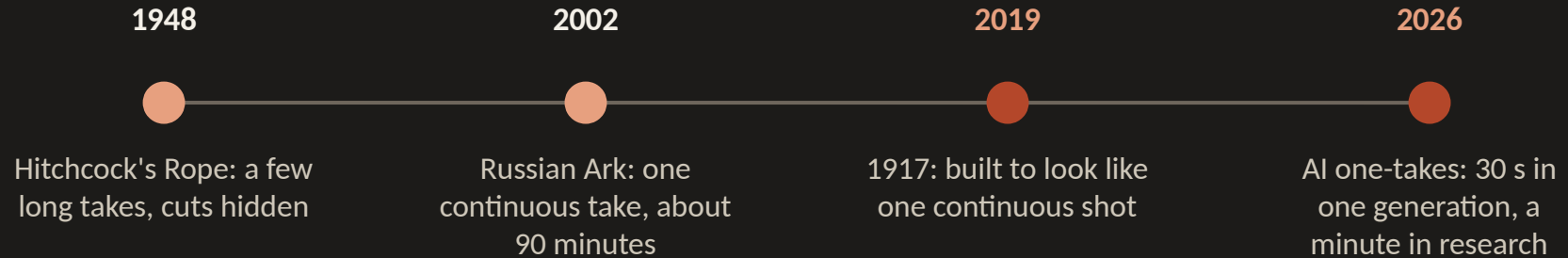
Stanford taught this in May

CS348K spent a lecture on how distillation and caching made autoregressive video real-time.

Coming in Week 10

Interactive world models and explorable art (Nov 3) start from exactly this idea.

The long take is an old dream



What 30 seconds unlocks: blocking

Performance blocking

Where actors stand and move across a whole scene, not one gesture.

Camera movement

A move with a start, a middle and an end, inside one take.

Green-screen and white-model control

Bring in your own blocked-out scene or plates to steer the result.

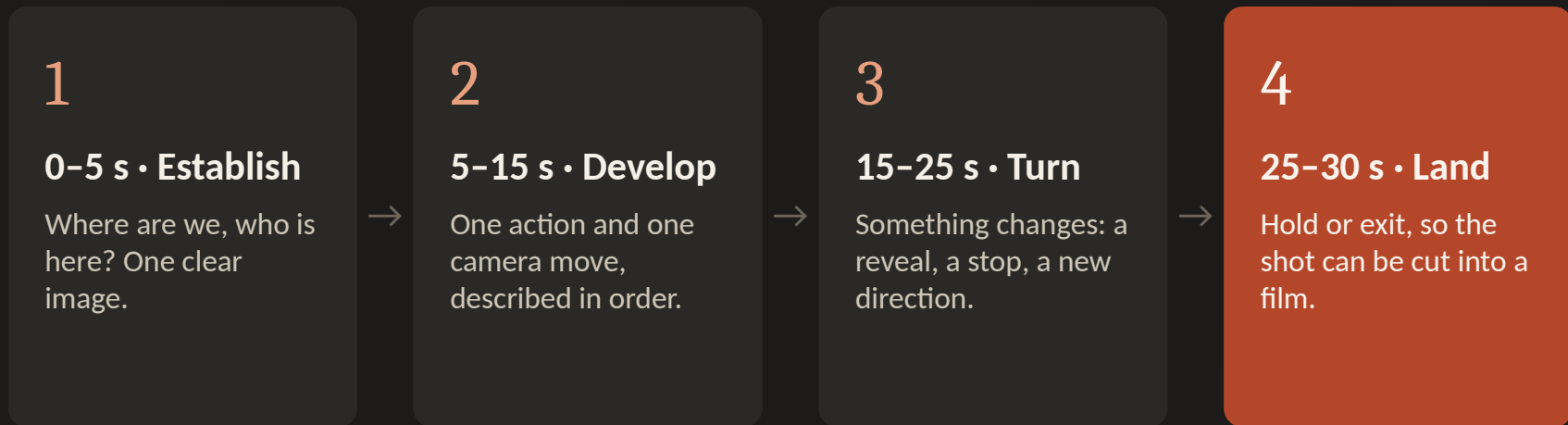
Extend twice

Continue a 30-second generation up to two more times.

▶ [Seedance 2.5 official page \(feature videos\)](https://seed.bytedance.com/en/seedance2_5)

seed.bytedance.com/en/seedance2_5

Useful: a beat sheet for a 30-second generation



A starting template, not a rule. Write the prompt in time order, one subject, one camera idea.

FOR THE ROOM

If one generation is a whole 30-second scene,
what is the director's job now?

And what did the editor just lose?

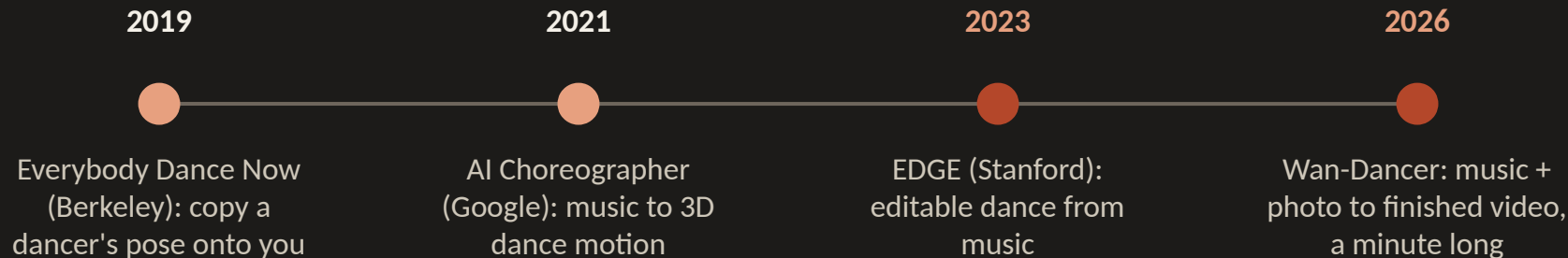
STORY 2 OF 4

02

Music becomes the choreographer

A song, one photo, a minute of dance

AI dance has a short, fast history



Under the hood: where dance models learn

AIST++ (Google, 2021)

5.2 hours of 3D dance, 1,408 sequences, 10 genres, filmed from 9 cameras. Small, clean, labelled.

Web video

Huge and varied, but noisy: no 3D, mixed camera work, and unclear consent.

Skeleton models

Trained on clean 3D motion like AIST++, so outputs are joints, not pictures.

Pixel models

Need far more video, so they lean on large web-scale data and a video foundation model.

Wan-Dancer: song in, dance out

A music track

One photo of the dancer

A style word



A 1-minute+ dance video, 720p / 30 fps,
on the beat

- Released Jul 13, 2026 by Tongyi Lab (Alibaba) as open weights, Apache 2.0
- Five styles: Chinese classical, K-pop, street, tap, Latin

 [Play: Wan-Dancer demos \(project page\)](https://github.com/humanaigc/wan-dancer-project)

[humanaigc.github.io/wan-dancer-project](https://github.com/humanaigc/wan-dancer-project)

[Paper \(arXiv 2607.09581\) ↗](https://arxiv.org/abs/2607.09581)

arxiv.org/abs/2607.09581

Skeleton then, pixels now

Before: music → skeleton

- Output is 3D joint motion
- You still need a character, costume, lighting and a renderer
- Great for games and animation pipelines

Now: music + photo → video

- Output is the finished image
- Look, light and camera come from the model
- Fast to try, harder to edit afterwards

The trade-off: control over every joint, or a finished shot in one step.

One dancer, many looks; one song, many dancers



- Swap the photo, keep the song: the same dance, a different dancer
- Swap the song, keep the photo: the same dancer, a new dance
- Earlier music-to-dance models fell apart after about 20 seconds

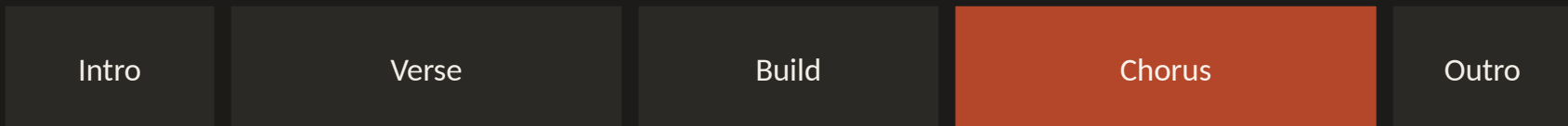
► **Play: compare with earlier models (\$6)**

humanaigc.github.io/wan-dancer-project

Reference images from the Wan-Dancer project page

The trick: plan the song, then the steps

Global level: the whole song



↓ plans

Local level: the moves

Renders each passage's steps, beat by beat

This is how a choreographer works: structure first, steps second.

Under the hood: two ways to stay coherent for a minute

One chunk at a time (Helios)

- Each chunk conditions on a compressed past
- Good for open-ended, interactive video
- Structure emerges; nothing plans the ending

Plan, then render (Wan-Dancer)

- A global model handles the long-range structure
- Local models render detail for each passage
- Good when the structure is known in advance: a song

Language models face the same choice: write word by word, or outline first. Music gives Wan-Dancer the outline for free.

What the model has to hear

Beat

Where the accents fall: a step, a hit, a tap must land on them.

Tempo

How fast the beats come: the same move looks different at 90 and 140 bpm.

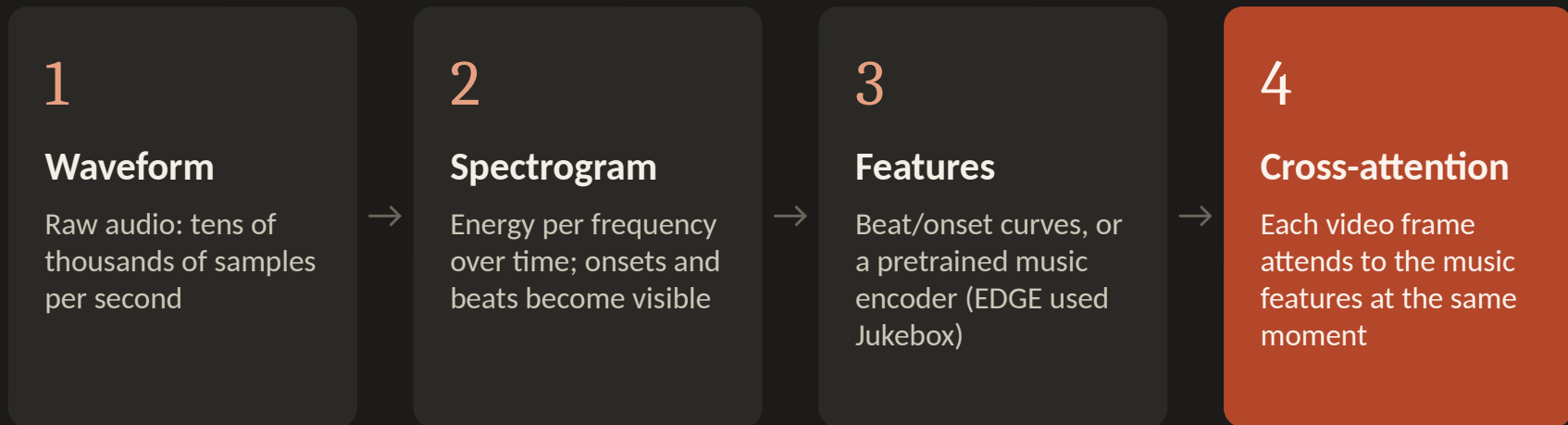
Phrases and sections

Eight-counts, verses, choruses: where a movement idea starts and ends.

Energy

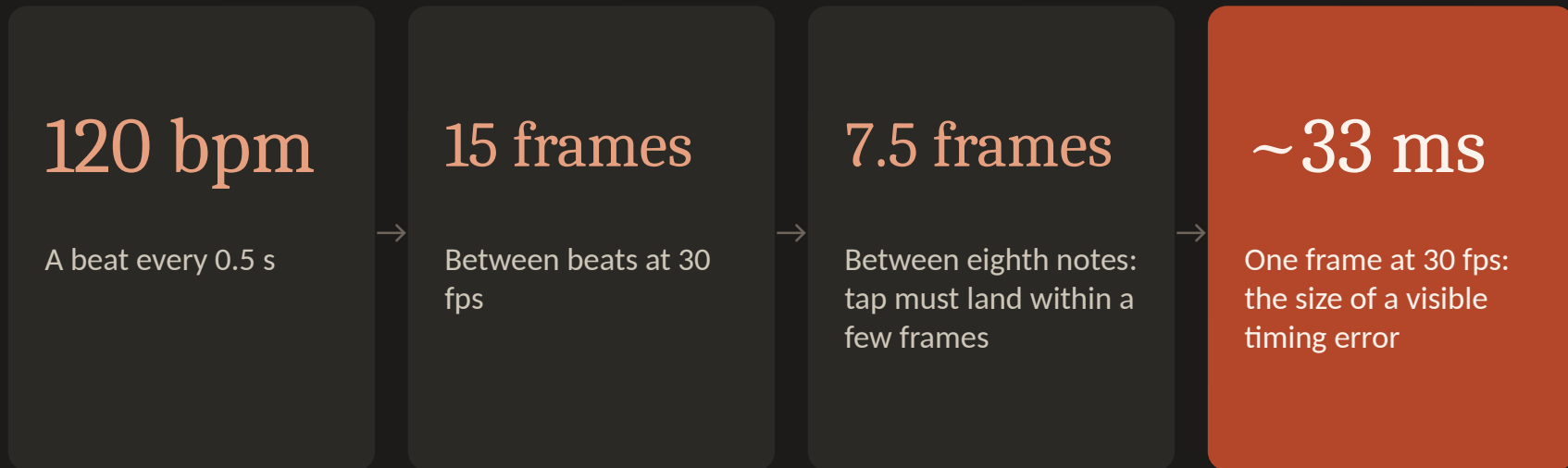
Quiet to loud, sparse to dense: when to hold still and when to explode.

Under the hood: how a model hears music



Same conditioning idea as text prompts in Weeks 2–3, with time-aligned audio instead of words.

Under the hood: beats and frames



Why tap is the hardest test: its timing tolerance is a couple of frames.

Five styles, five tests

Chinese classical

Flow: does the movement stay continuous through the phrase?

K-pop

Precision: are the hits sharp and exactly on the beat?

Street

Groove: does the body bounce with the rhythm, not just pose?

Tap

Sound: do the feet land where the taps are heard?

Latin

Hips and rhythm: is the movement driven from the core?

Your verdict

Which style does the model handle worst, and why?

► **Play: all five styles (project page, section 1)**

humanaigc.github.io/wan-dancer-project

Machines in choreography are not new

Merce Cunningham • chance

Used chance procedures, such as tossing coins, to order movements and decide structure.

Cunningham • Trackers, 1991

The first dance he made partly with LifeForms, an early computer program for figures in motion.

Wayne McGregor • Living Archive, 2019

An AI tool trained on his 25-year archive suggests the next phrase; dancers translate it through their bodies.

[Living Archive \(Studio Wayne McGregor\) ↗](https://waynemcgregor.com/productions/living-archive)

waynemcgregor.com/productions/living-archive

Limits to keep in mind

Hardware

Open weights, but its release materials point to datacenter-class GPUs, not a laptop.

One dancer, one photo

The look comes from a single image; everything unseen in it is invented.

Whose moves?

Viewers cannot check whose dances the training videos contained.

Music has rights

The song that drives the dance is someone's work; credit it and check its licence.

Pitch a one-take

5 MINUTES • IN PAIRS

1

One song: name it, and say why

2

One photo: who or what dances?

3

One sentence: what happens in the 30 seconds?

Three pairs read theirs aloud. Keep your pitch: it could become Assignment 2.

FOR THE ROOM

Who is the choreographer here?

The musician? The model? The thousands of dancers in its training videos? The person who picked the song?

STORY 3 OF 4

03

Motion leaves the screen

A phone video becomes a robot's dance

Robots have danced before

2000s



HRP-2 learns a Japanese folk dance from motion capture

Dec 2020



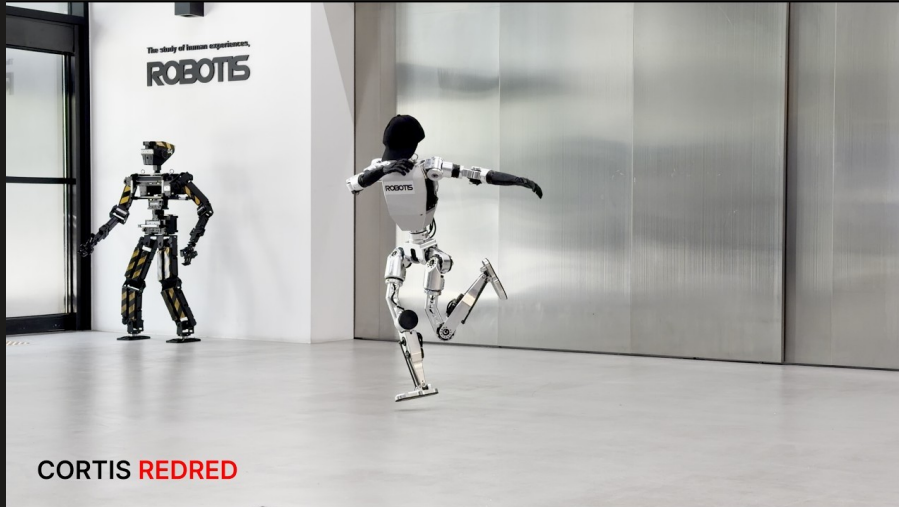
Boston Dynamics “Do You Love Me?”: hand-authored routine

Jun 2026



AI Sapiens learns a K-pop challenge from phone video

A humanoid learns a K-pop challenge from phone video



- ROBOTIS (Korea), open-source humanoid AI Sapiens, June 2026
- 1.3 m tall, 34 kg, 23 joints
- Learned from ordinary 2D video: no motion-capture studio

ROBOTIS AI Sapiens dancing the CORTIS “REDRED” challenge (ROBOTIS video thumbnail)

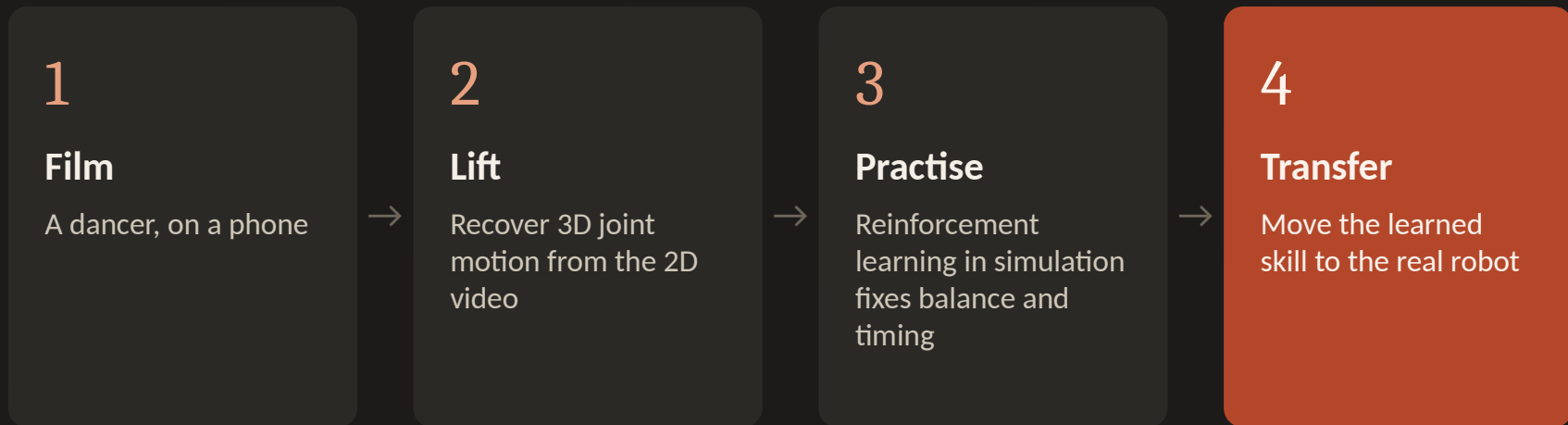
▶ [Play: ROBOTIS dance video \(YouTube\)](#)

www.youtube.com/watch?v=NeL2BmgcFJs

[News story \(Interesting Engineering\) ↗](#)

interestingengineering.com/ai-robotics/korean-humanoid-r...

Four steps from phone to robot



Every step is an idea from this course: pose from video, simulation, and a model of the world.

Under the hood: 3D motion from one camera

2D keypoints

Find joints in each frame: wrists, knees, ankles. Mature and fast.

Lift to 3D

Fit a body model such as SMPL (24 joints) so the 2D points make sense in 3D.

Hard parts

Depth is ambiguous from one camera; fast motion blurs; feet slide if ground contact is wrong.

Why not just copy the pose?

Different body

1.3 m, 34 kg, 23 joints: limbs and proportions don't match the dancer's.

Speed limits

Motors can't accelerate like muscles. A 2025 study copying a dancer's arms found the robot couldn't keep the rhythm.

Balance

A human shifts weight without thinking. Copy the pose exactly and the robot falls over.

[Study: robot arms copying a dancer \(2025\) ↗](#)

doaj.org/article/b55086ceb9bf4dd3b6b2893ef09cb7db

Under the hood: retargeting

Map the joints

Match human joints to robot joints; drop or merge those the robot doesn't have.

Scale and limit

Rescale limb lengths; clamp angles and speeds to what the motors allow.

Fix the contacts

Use inverse kinematics so hands and feet land where they should.

Under the hood: learning to imitate

Policy

A neural network that reads the robot's state and outputs motor commands, many times a second.

Tracking reward

Reward matching the reference: joint angles, velocities, hand and foot positions, body root.

Stay alive

Penalise falling, jerky motion and wasted energy, so it dances and survives.

Where to learn it

DeepMimic (SIGGRAPH 2018) set the recipe; Berkeley's CS 285 teaches the RL behind it.

The sim-to-real gap

Practice in a copy of the world

The robot's twin trains in NVIDIA Isaac Sim, falling thousands of times at no cost.

But the copy is imperfect

Friction, motor delay and flex differ between simulation and the real floor.

Close the gap in hardware

ROBOTIS designed its DYNAMIXEL-Q actuators so simulated joints behave more like real ones.

A course idea in disguise

Simulation is a world model. We'll meet video world models in Week 10.

Under the hood: domain randomization

Randomize the world

Each practice run gets different friction, masses, motor delays and pushes.

Learn what always works

A policy that survives a thousand slightly wrong worlds treats reality as one more.

The art connection

Rehearsing on different stages and floors does the same for human dancers.

Why dance challenges are perfect training data

Short

Fifteen to thirty seconds: a complete, learnable unit.

Repeated

Thousands of people film the same routine: many examples of one choreography.

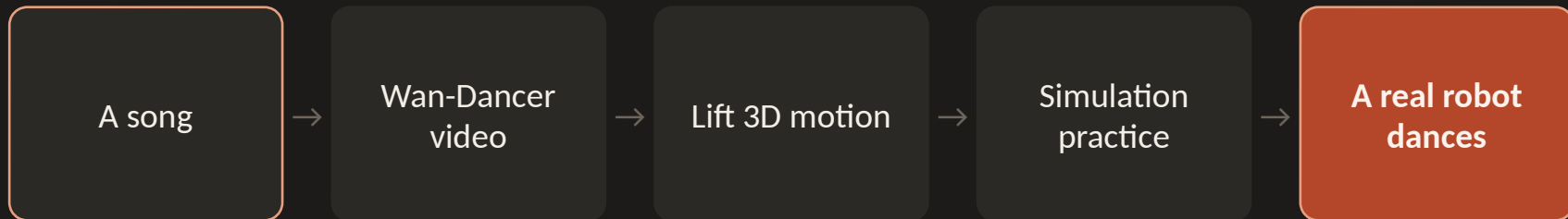
On the beat

The music marks exactly when each move should happen.

Full body, facing the camera

Made to be watched and copied, which also makes them easy to read.

A thought experiment



Speculation: as far as public reports go, nobody has shown this end to end. But every piece now exists.

Video is becoming training data for bodies

For robots

Human videos teach skills without motion capture: dance today, chores tomorrow.

For world models

Video generators that predict what happens next can act as simulators (Week 10).

For artists

Your choreography video may have two audiences: people, and machines that learn from it.

FOR THE ROOM

If a generated video can teach a real body to move, is that video still “fake”?

What is a dance video for, once machines learn from it?

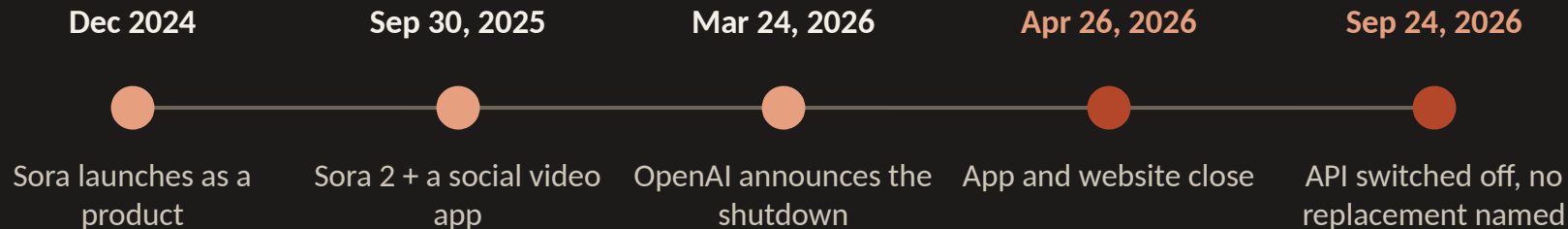
STORY 4 OF 4

04

Sora is gone

When your tool disappears

Eighteen months as a product



[OpenAI: What to know about the Sora discontinuation ↗](https://help.openai.com/en/articles/20001152-what-to-know-about-...)

help.openai.com/en/articles/20001152-what-to-know-about-...

Open or closed, this year

Model	Status	Can you still run it in 5 years?
Helios	Open weights	Yes, if you keep a copy
Wan-Dancer • Wan 2.2	Open weights (Apache 2.0)	Yes, if you keep a copy
Seedance 2.5 • Wan 3.0 • Kling 4.0	Closed; app or API only	Only if the company keeps it running
Sora 2	Discontinued	No

Under the hood: can you run an open model?



Weights only; activations, the VAE and text encoders need more memory on top.

This has happened before: Flash

Dec 31, 2020

Adobe ended support for Flash Player. Two decades of web animation and games stopped playing in browsers.

Rescue by emulation

The Internet Archive keeps thousands of Flash works playable with an emulator, Ruffle.

Art institutions

Rhizome's ArtBase has preserved born-digital art since 1999. Preservation is a practice, not an accident.

Last fall, a top syllabus was built on Sora

MIT's Fall 2025 motion-graphics course listed Sora as its video tool. Some of last year's student projects depend on a tool that no longer exists.

Keep your sources, not just your outputs

- ✓ Prompts, word for word
- ✓ Seeds and settings
- ✓ Reference images, clips and songs (with credit)
- ✓ Raw generations, before editing
- ✓ Edit project files
- ✓ Model name and version, and the date

Why open weights matter

Open-weight models (Wan 2.2, Helios, Wan-Dancer) can still be run years from now. Closed apps can't.

A project folder that survives

README.md

Model, version, date, settings, and what you were trying to do.

prompts/

Every prompt you kept, word for word, with its seed.

refs/

Images, clips and songs you used, with credits and licences.

raw/

Unedited generations, including the failures worth remembering.

edit/

The editing project and the exported master.

Start today

Use this structure for Assignment 2.

Useful: pin your versions

Name the exact model

Write the model ID and version, not just “Wan”: e.g. Wan-AI/Wan-Dancer-14B.

Pin the revision

On Hugging Face, record the commit hash of the weights you downloaded.

Keep the licence

Save the licence file next to the weights: Apache 2.0 lets you keep and share them.

Write the settings

Seed, steps, guidance, resolution, frame rate: everything needed to make it again.

FOR THE ROOM

What does an AI artwork need in order to outlive the tool that made it?

Think of a film, a painting, a game. What survives, and why?

Four stories, one line each

01 • Minutes

The unit of AI video moved from the shot to the scene, and toward real time.

02 • Music

A song and a photo now give you a minute of dance, planned like a choreographer would.

03 • Machines

Phone video can teach a real robot to dance; video now has machine audiences.

04 • Loss

Tools can vanish. Keep your sources; prefer what you can still run.

DEBATE • 10 MINUTES

“A generated dance that a robot can learn is a real dance.”

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

Debate starters

For: it is a real dance

- Choreography is a structure of movement in time; the model plans one
- A real body performs it, with real risk of falling
- Dance has always borrowed from tools: chance, software, video

Against: it is not

- No one intended it; a model averaged other people's dances
- The robot executes, it doesn't express
- Without a dancer's choice, it's motion, not dance

Use one example from tonight in your minute.

Useful: try it this week

What	Where	What you need
Helios, real-time long video	Hugging Face demo: BestWishYsh/Helios-14B-RealTime	A browser
Wan-Dancer, music to dance	ModelScope demo: Wan-AI/Wan-Dancer	A song you may use + one photo
Seedance 2.5, 30-s one-take	ByteDance's apps and API	An account; use the beat sheet
AI Sapiens, video to robot	docs.robotis.com (AI Sapiens)	Curiosity; read the pipeline

▶ [Helios demo \(Hugging Face\)](https://huggingface.co/spaces/BestWishYsh/Helios-14B-RealTime)

huggingface.co/spaces/BestWishYsh/Helios-14B-RealTime

This week

READING 1 • TECHNIQUE

[Helios: real-time long video generation](#)

READING 2 • ART + TECH

[Wan-Dancer project page](#)

READING 3 • CULTURE

[What happened to Sora?](#)

[Quiz 07 • two technique, two art, one reflection • hyang.org/amcc5160/quiz-07](https://hyang.org/amcc5160/quiz-07)

Optional technical reading: Self Forcing (NeurIPS 2025) • Assignment 2 is due Oct 31 • Next week, Lecture 08: native audio-video, performance and real-time characters

All demo links in one place

STORY 1

- ▶ [Seedance 2.5 official page](#)
- ▶ [Helios feature video \(YouTube\)](#)
- ▶ [Helios project page](#)
- ▶ [Self Forcing paper \(NeurIPS 2025\)](#)

STORY 3

- ▶ [ROBOTIS: Learning from Videos \(YouTube\)](#)
- ▶ [AI Sapiens documentation](#)
- ▶ [Study: robot arms copying a dancer](#)

STORY 2

- ▶ [Wan-Dancer project page](#)
- ▶ [Wan-Dancer weights \(Hugging Face\)](#)
- ▶ [Living Archive \(Wayne McGregor\)](#)
- ▶ [Wan-Dancer demo \(ModelScope\)](#)

STORY 4

- ▶ [OpenAI: Sora discontinuation](#)

Music, minutes, and machines

Motion got longer, found a new choreographer, left the screen, and lost a tool. Keep your sources.

Sources: Seedance 2.5 (ByteDance Seed); Helios (Peking University, ByteDance, arXiv 2603.04379); Wan-Dancer (Tongyi Lab, arXiv 2607.09581); ROBOTIS AI Sapiens; OpenAI Help Center; reviews by Pinggy, Atlas Cloud and Unifically for Wan 3.0 and Kling 4.0 specs. Images: Helios feature video, Wan-Dancer project page, ROBOTIS video thumbnail, used for teaching.