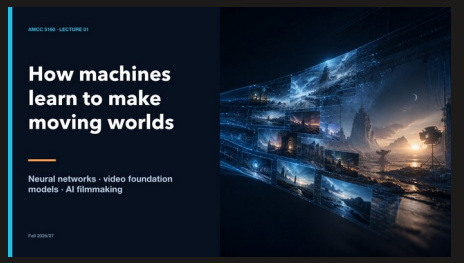
AMCC 5160 · LECTURE 11 · 10 NOVEMBER 2026

The Semester *in One Evening*

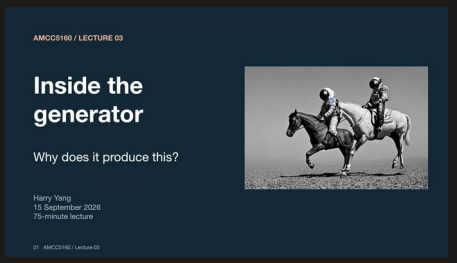
Review of Lectures 01–10: the key technical ideas, with the original slides

Review session · Fall 2026

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



Lecture 01



Lecture 03



Lecture 05



Lecture 07



Lecture 08

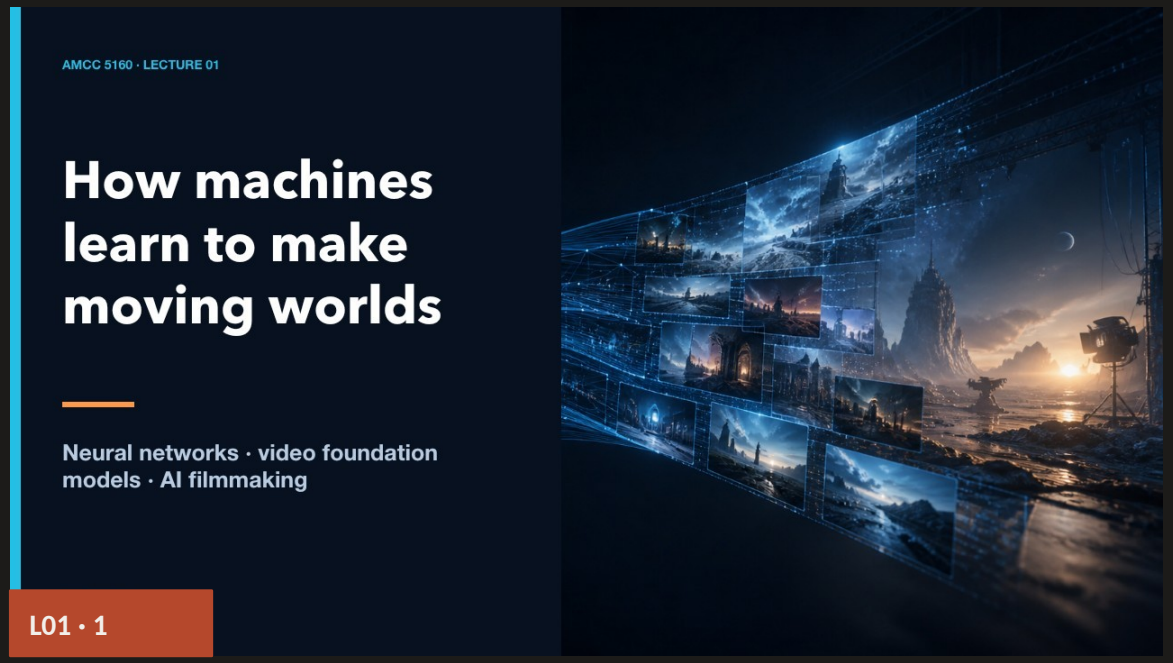


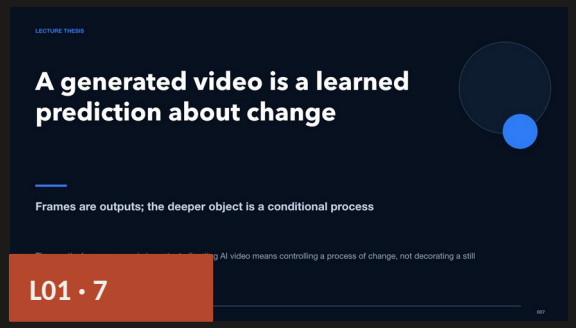
Lecture 09

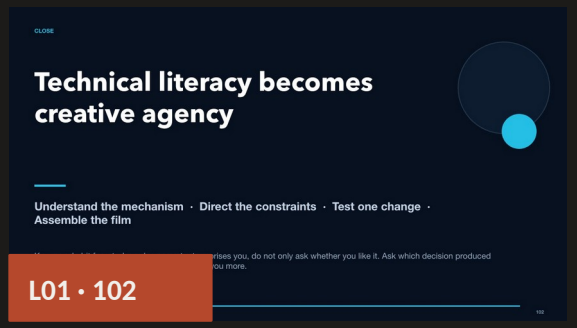
Ten Tuesdays on one line

01 · Sep 1 **Neural networks → AI filmmaking** *A generated video is a learned prediction about change*	02 · Sep 8 **Generative editing** *Write the edit contract: change, keep, adapt*	03 · Sep 15 **Inside the generator** *The model predicts; the sampler takes a step*	04 · Sep 22 **Video generation + art filmmaking** *Two kinds of time; intention comes first*	05 · Sep 29 **Making the model yours** *Whose model is it?*
06 · Oct 6 **Quantization (TA talk)** *Store less: fewer bits per number*	07 · Oct 13 **Music, minutes, machines** *The unit is now the scene; keep your sources*	08 · Oct 20 **AI video that talks** *One model, one clock; who is the performer?*	09 · Oct 27 **Worlds, 3D, robots** *Add an action and video becomes a place*	10 · Nov 3 **Critique + provenance** *Is it AI? good? fair? allowed?*

Weeks 12–13: final presentations · Nov 17 and Nov 24







01

LECTURE 01 · 1 SEPTEMBER 2026

Neural networks to AI filmmaking

What is a generated video, really?

WE COVERED

- Neurons, CNNs, attention and transformers
- VAEs, diffusion and flow matching
- Video as a temporal system; Seedance and MiniMax H3
- A five-gate AI filmmaking workflow

Every generator in four stages

Data

Examples, captions, filters

Lecture 04: filtering trades coverage for quality

→

Representation

VAE latents, patch tokens

Lecture 01: compress first, then create

→

Model

Transformer with attention

Lecture 03: predicts noise or velocity

→

Sampling

Seed, guidance, steps, sampler

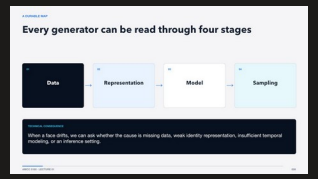
Lecture 03: the sampler takes the step

DIAGNOSE

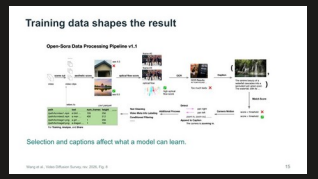
A face drifts. Which stage would you test first, and how?

REVISIT

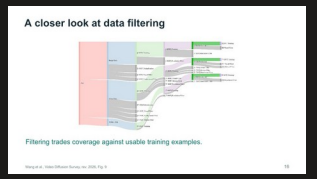
Click to open



L01 · 8



L04 · 15



L04 · 16

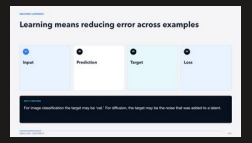
From one neuron to a video model, in five steps

1	→	2	→	3	→	4	→	5
Neuron		**CNN**		**Attention**		**VAE latent**		**Diffusion · flow**
Weighted sum, bias, nonlinearity; learn by reducing error		One learned filter scanned across the image		Every token chooses what matters; RoPE adds order		Compress first, then learn to create		Learn the path from noise to data
L01 · 11–12		L01 · 16		L01 · 23–29		L01 · 32–35		L01 · 36–41

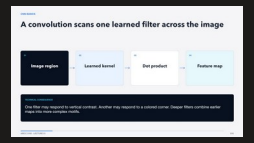
Every model you met later rearranges these five parts.

REVISIT

Click to open



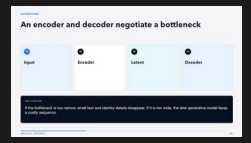
L01 · 11



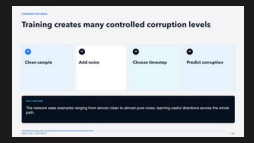
L01 · 16



L01 · 26



L01 · 33



L01 · 36

Compress first, then create

240

frames in 10 s at 24 fps

Lecture 01

16× · 4×

spatial · temporal compression in H3

Lecture 01

6× → ~36×

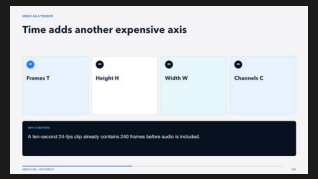
longer clip → attention comparisons

Lecture 07

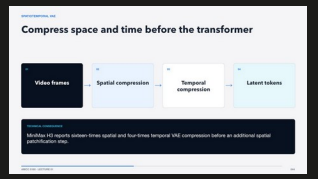
Compression is a design requirement, not an optimisation. It decides which details can survive.

REVISIT

Click to open



L01 · 43

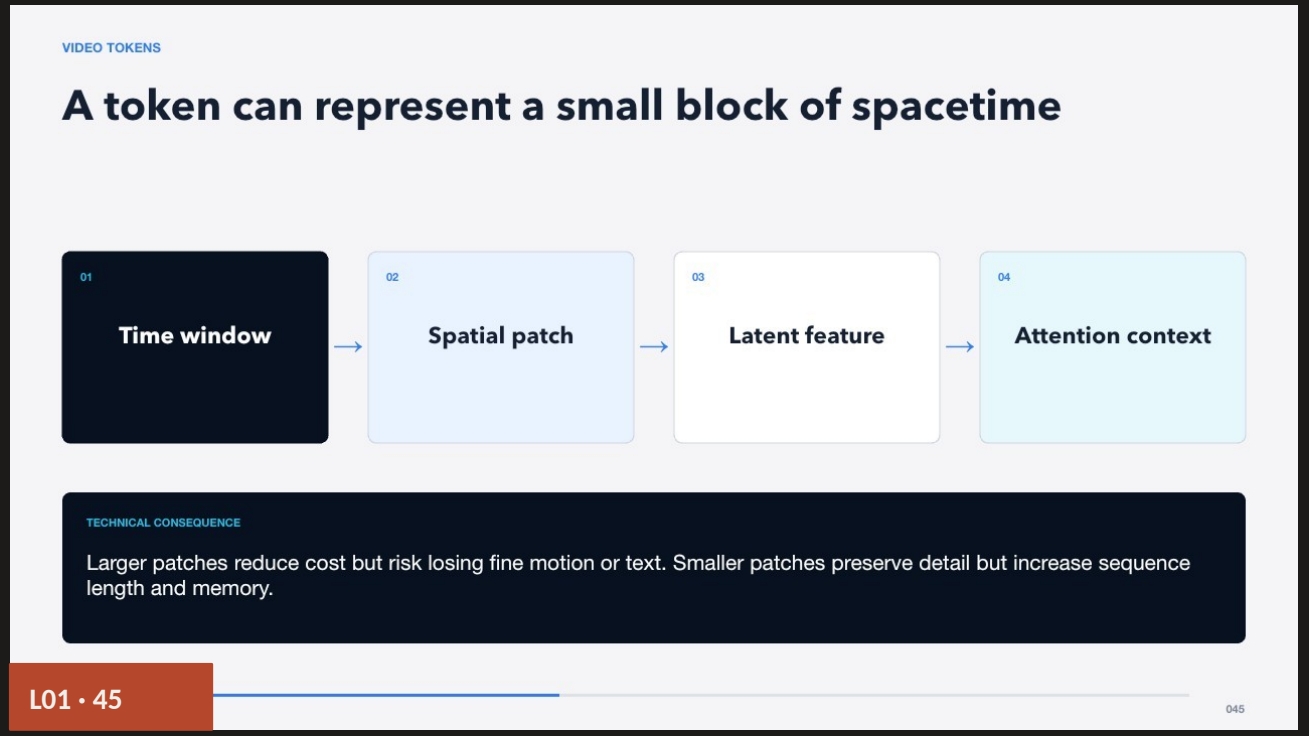


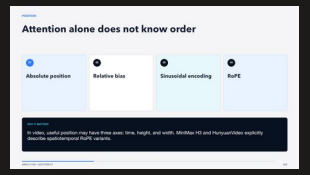
L01 · 44



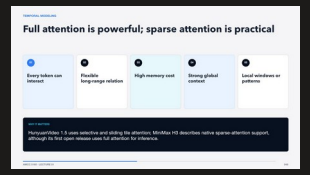
L07 · 9

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





L01 · 29



L01 · 46

A token is a block of space-time

Bigger patches: cheaper, but lose fine motion and text.

Attention needs coordinates

RoPE rotates Q and K; video needs time, height and width.

Who may attend to whom

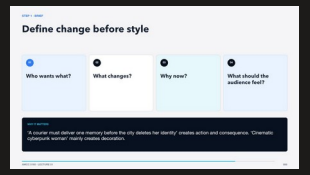
Full attention is quadratic; sparse, sliding and block attention are practical.

ASK YOURSELF

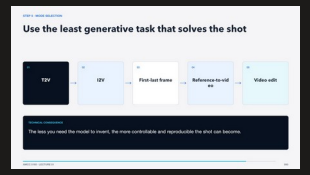
For any model you used: what is a token, where is it, and who can it see?

A film is a sequence of decisions, not one generation





L01 · 86



L01 · 90

Five gates

Brief → world bible → animatic → shots → post. Each gate leaves an artifact you can inspect.

Define change before style

Agent, goal, change, stakes. “Cinematic cyberpunk woman” is decoration, not an event.

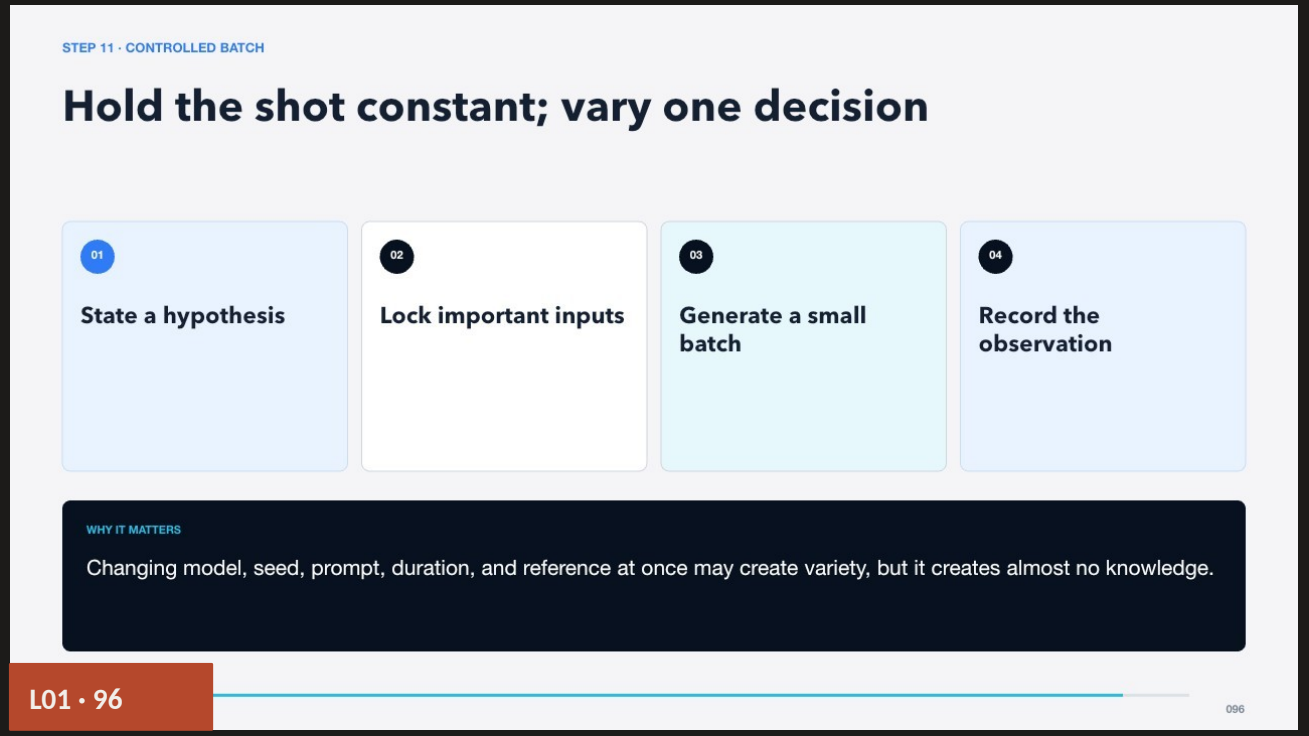
Use the least generative task

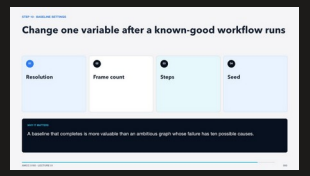
T2V invents; I2V keeps a frame; first-last controls endpoints; V2V keeps blocking.

ASK YOURSELF

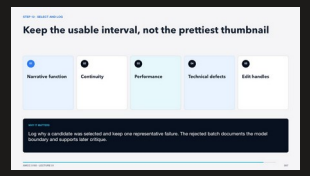
For your final project: which gate did you skip, and what did it cost you?

Hold the shot constant; vary one decision

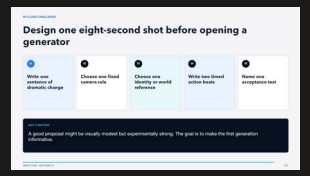




L01 · 95



L01 · 97



L01 · 101

Known-good first

Run the official template once; record frames, fps, steps, guidance, sampler, seed.

One variable at a time

3–6 candidates; write observations before choosing. Many changes = no knowledge.

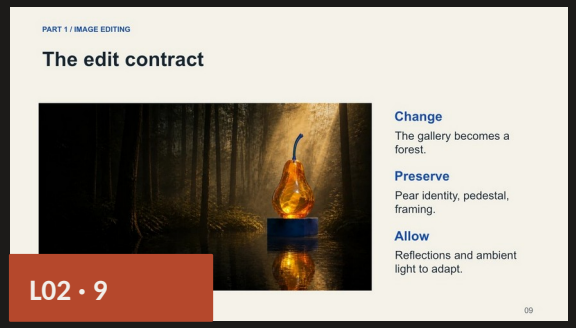
Keep the usable interval

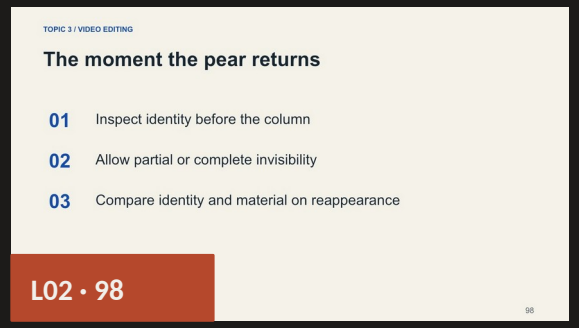
Full speed, in/out points, one representative failure. Edit first, repair later.

ASK YOURSELF

Design one eight-second shot: one change, a camera constraint, two beats, one acceptance test.







02

LECTURE 02 · 8 SEPTEMBER 2026

Generative editing

When you ask for a change, what should stay the same?

WE COVERED

- The edit contract and its three columns
- Controls: text, mask, ControlNet, references, LoRA
- Art direction for AFTER RAIN
- Video editing: survive the next frame

The edit contract: change, keep, adapt

Change

What the brief asks for

The environment: gallery → forest

Keep

Invariants: semantic or exact?

The pear's identity, pedestal, framing

Adapt

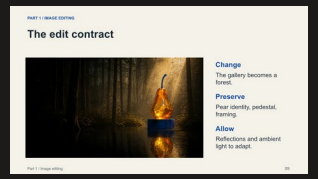
Consequences that must follow

Light, shadows, reflections

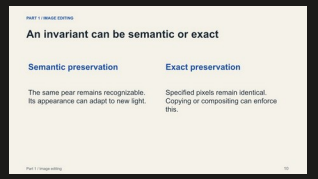
“Keep it the same” can mean recognizable or pixel-identical. Only copying or compositing can promise pixel-identical.

REVISIT

Click to open



L02 · 9



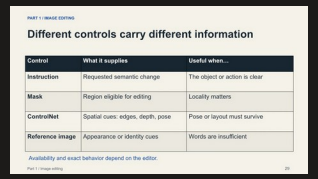
L02 · 10

Choose a control by the information that is missing

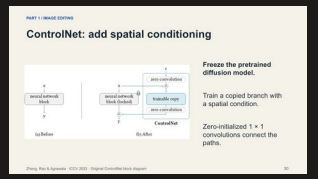
Control	It supplies…	Watch out for…
Text prompt	What should change	A condition, not a hard constraint
Mask	Where change is allowed	Seams and reflections are not solved by the mask
ControlNet: edges · depth · pose	Shape and structure to follow	Says nothing about material or identity
Reference image · IP-Adapter	Appearance words can't pin down	Instant, but less faithful than training
LoRA · DreamBooth	A concept that must recur	Did it learn the subject, or the setting?

REVISIT

Click to open



L02 · 29

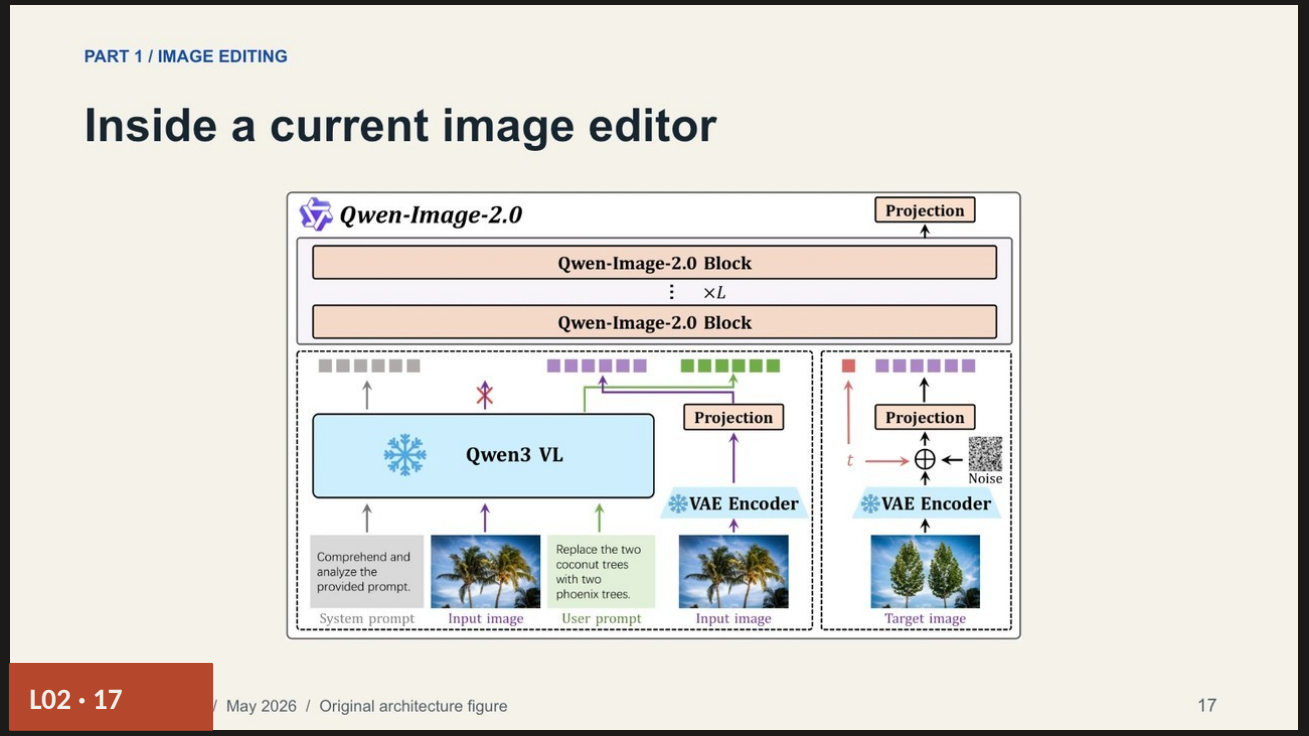


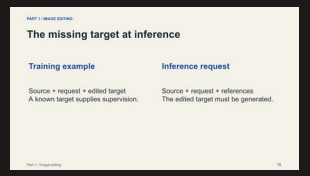
L02 · 30



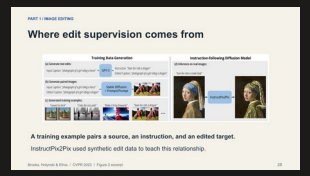
L05 · 64

Inside an image editor: the target is missing at inference

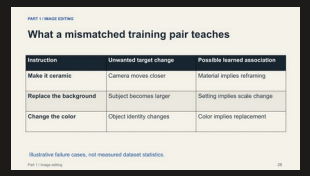




L02 · 18



L02 · 25



L02 · 26

The source enters twice

Semantic features (what is “it”?) and VAE latents (this particular pear).

The target is missing at inference

A reference conditions one generation; it does not teach weights.

Supervision teaches side effects

If targets also move the camera, the model learns to move the camera.

ASK YOURSELF

When your edit failed: did the model misunderstand, or lose visual information?

Count what the model actually receives

Image latents

$$1024 \times 1024 \times 3 \rightarrow 64 \times 64 \times 64$$

3,145,728 → 262,144 (÷12, not ÷16)

Channels change too: count values, not one axis

Video latents

$$45 = 4k + 1 \rightarrow k = 11$$

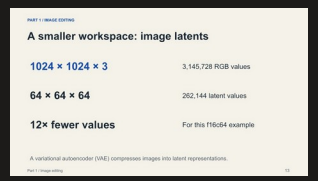
12 latent frames → a 3 × 4 grid

The first frame is special; 45 ÷ 4 misses it

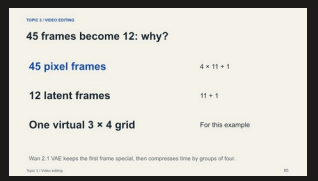
InstructPix2Pix: two scales, image and text. At s = 1 both differences cancel to the full prediction.

REVISIT

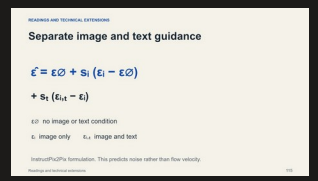
Click to open



L02 · 13

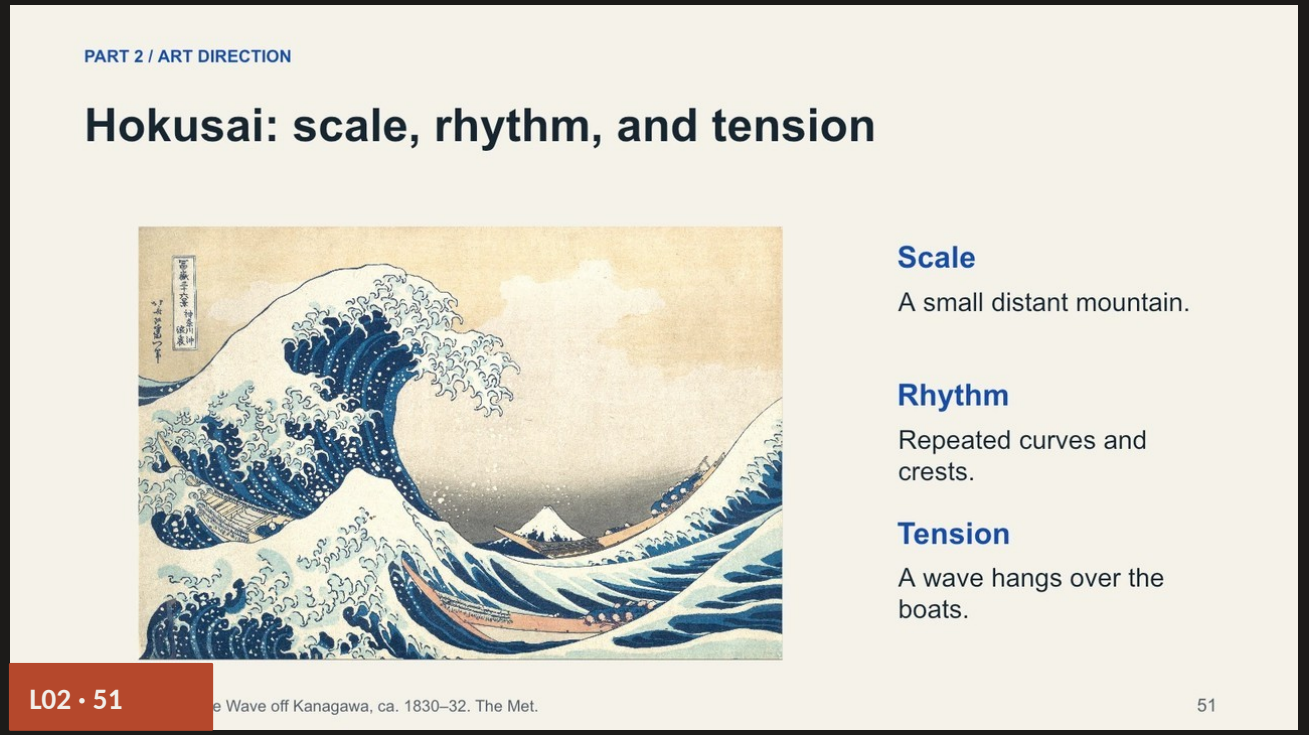


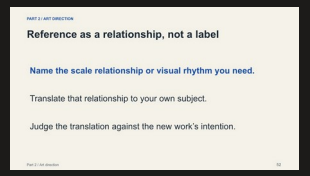
L02 · 85



L02 · 115

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





L02 · 52



L02 · 57



L02 · 54

Make the word visible

"Fragile" becomes scale, restrained light, an unstable reflection: hypotheses to test.

Borrow a relationship

Hokusai: a small form under a large curve. Not a style label.

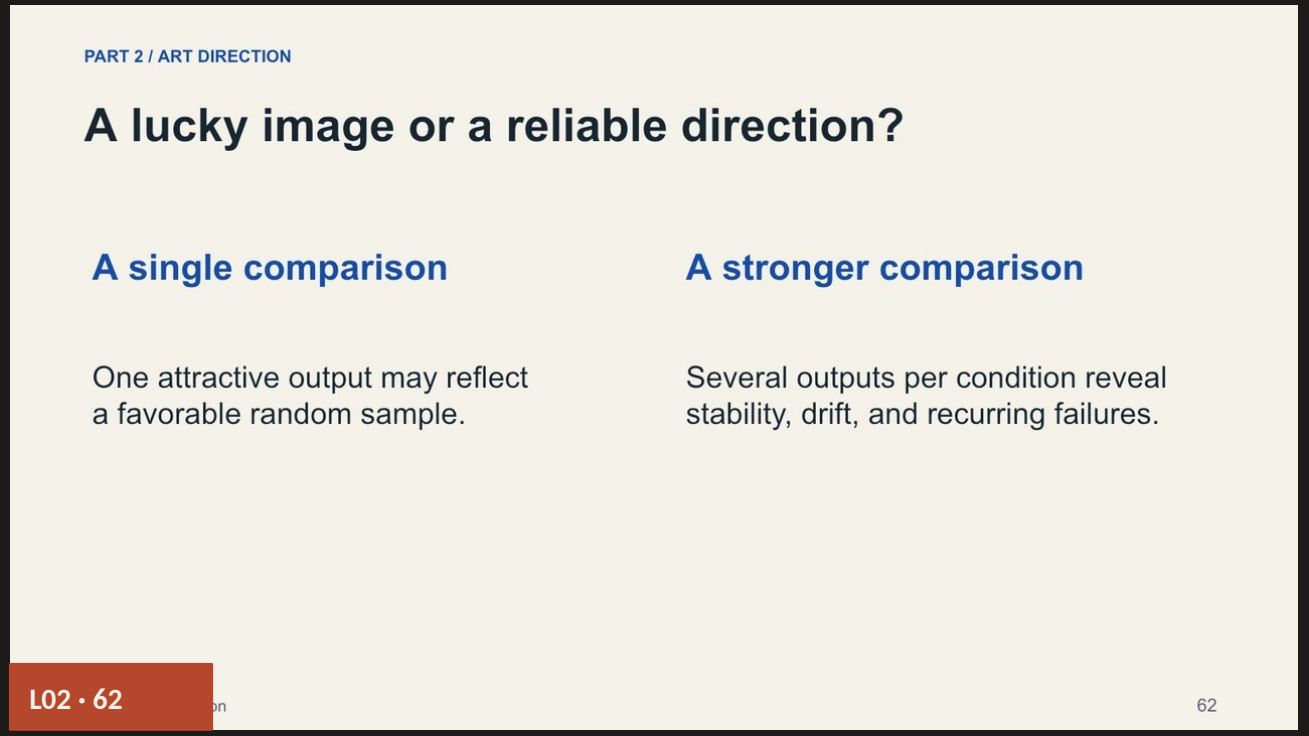
Give each reference a role

Subject, composition, material. Remove one and see what disappears.

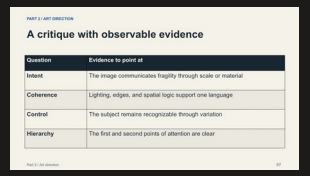
ASK YOURSELF

Which reference in your project has no assigned role?

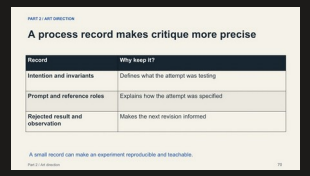
A lucky image or a reliable direction?



L02 · 62



L02 · 67



L02 · 70

Lucky or reliable?

Several outputs per condition. “Would I exhibit this?” ≠ “Can I build a series?”

Feature → effect → revision

“The crisp reflection makes it photographic; simplify its shape.”

Keep a process record

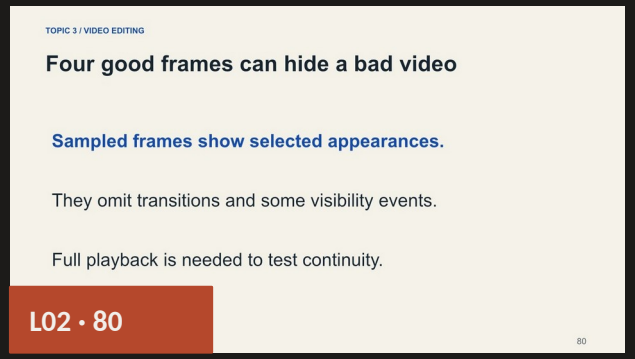
Intention, invariants, reference roles, one rejected result and why.

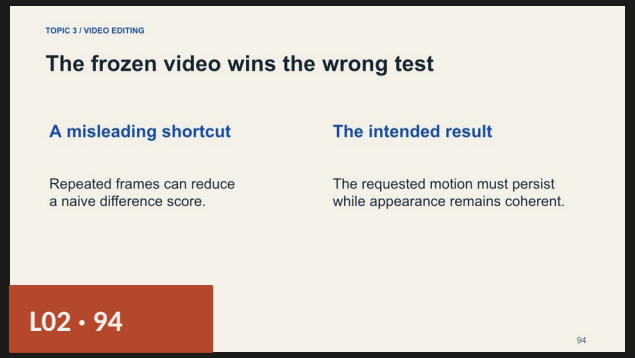
ASK YOURSELF

Rewrite one comment you received this term as feature → effect → revision.

A video edit must survive the next frame







1 **Consistency is not stillness**

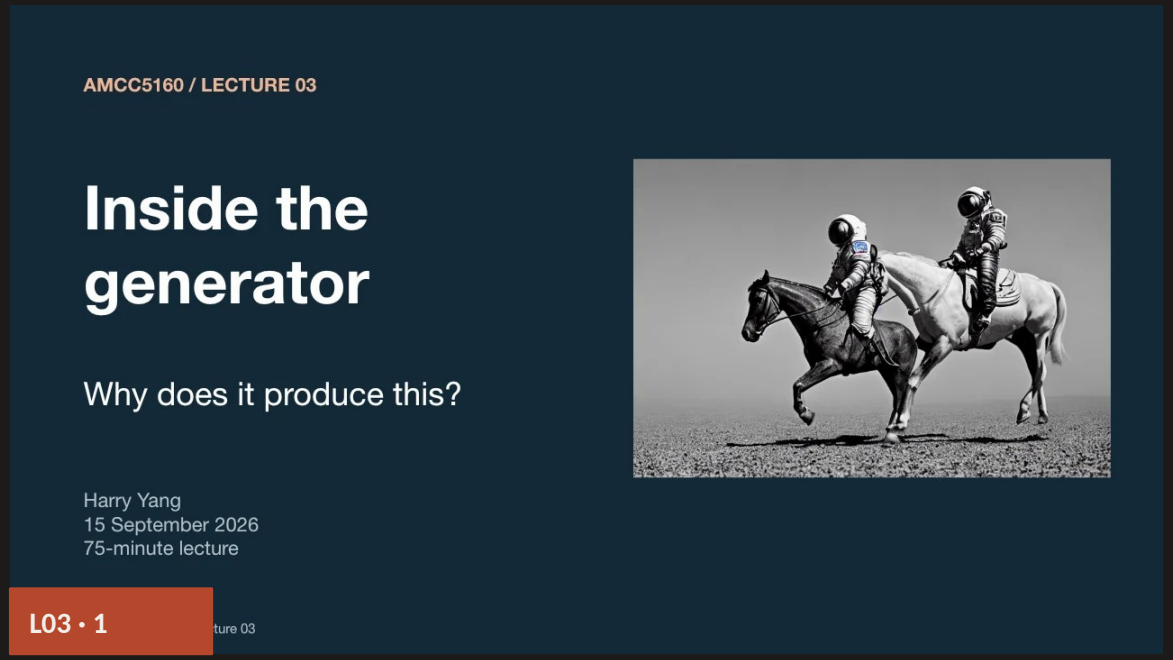
Change should be explained by the scene: pose, light, occlusion. (L02 · 76)

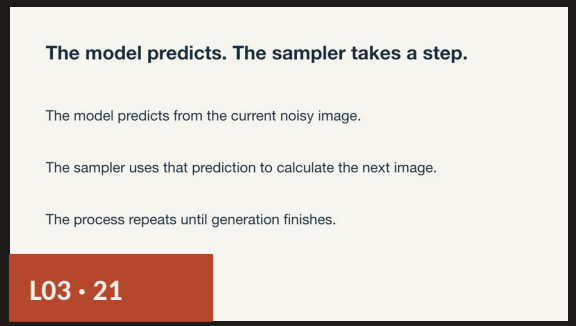
2 **Four good frames can hide a bad video**

Contact sheets locate problems; full playback finds them. (L02 · 80)

3 **The frozen video wins the wrong test**

Raw frame difference rewards a clip that never moves. (L02 · 93–94)







03

LECTURE 03 · 15 SEPTEMBER 2026

Inside the generator

Why does the generator produce this picture?

WE COVERED

- Noise, denoising, VAEs and flow matching
- CLIP, attention and transformers
- Seeds, guidance, steps; LoRA and ControlNet
- First look at Astra, world models and VLAs

Close is not correct: CLIP, relations and continuity

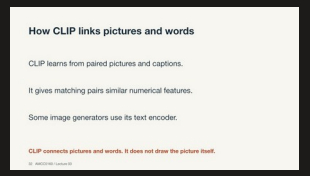
A similar-looking image can still be wrong

The picture may show a station and a traveler.

It may still put the suitcase on the wrong side.

Check specific details, not just the overall resemblance.

L03 · 33 Lecture 03



L03 · 32



L03 · 51

CLIP learns associations

Captions name the person and skip the railing, so relations are weakly learned.

Check relations, not nouns

The suitcase on the wrong side; two objects become three.

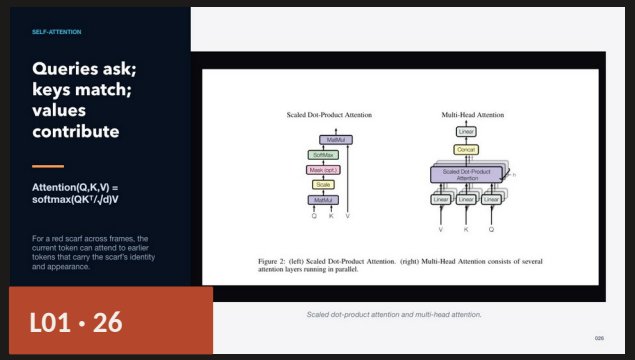
A character sheet, not a seed

Silhouette, surface details, ownership: which hand holds the prop.

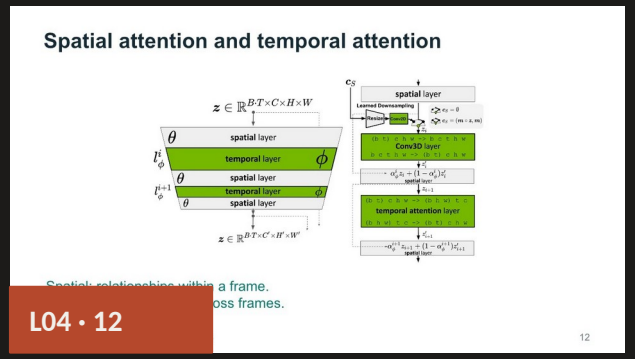
ASK YOURSELF

Write three relation checks for your final project's key frame.

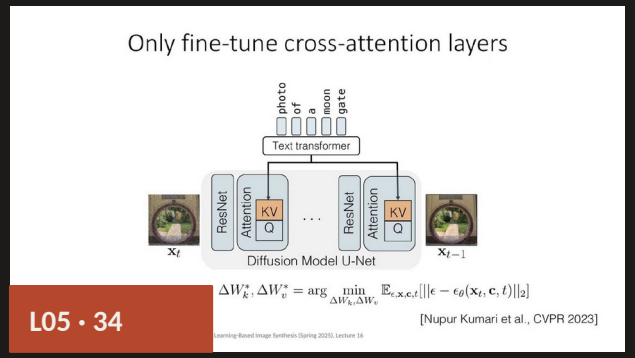
Attention: every token chooses what matters



L01 · 26



L04 · 12



L05 · 34

1 **Queries ask, keys match, values contribute**

Softmax weights sum to one. Many heads learn different relations. (L01, L03)

2 **Cross-attention is where words meet pixels**

Custom Diffusion fine-tunes only its K and V, and still learns a concept. (L05)

3 **Spatial within a frame, temporal across frames**

Connecting features is not the same as understanding the scene. (L04)

Learn to remove noise, then learn how to move

Diffusion: add noise

$$x_t = a \cdot x_0 + b \cdot \varepsilon$$

$0.6 \times 0.8 + 0.8 \times (-0.2) = 0.32$

The network learns to predict ε, at every noise level

Flow matching: learn to move

$$x_t = (1 - t) \cdot \varepsilon + t \cdot x$$

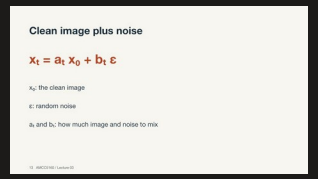
velocity = x − ε

At generation time there is no target; follow the learned field

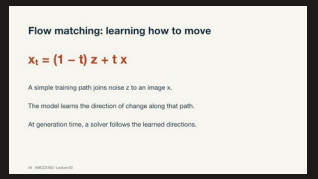
The model predicts. The sampler takes the step.

REVISIT

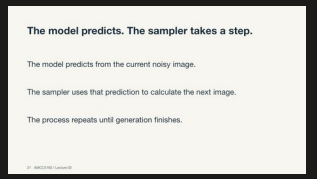
Click to open



L03 · 13



L03 · 44



L03 · 21

Three knobs, three different jobs

Seed

An address, not an instruction

Save model, prompt, references and settings, not just the number.

L03 · 48–49

Guidance

$\varepsilon = u + s \cdot (c - u)$

$u = 2$, $c = 3$, $s = 4 \rightarrow 6$. It extrapolates: stronger subject, burnt colours, less variety.

L03 · 52–55

Steps

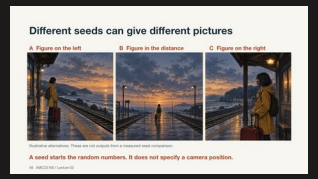
Accuracy of the path, not of the map

More steps ≠ more correct. A distilled 1–4 step model is a different trained system.

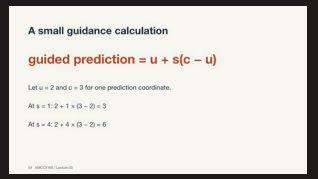
L03 · 60–64

REVISIT

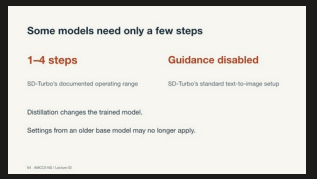
Click to open



L03 · 48



L03 · 54



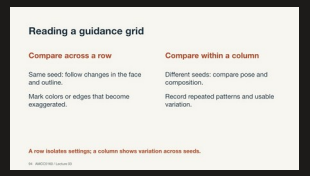
L03 · 64

How to make a fair comparison

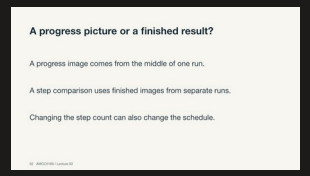
How to make a fair comparison

Part of the test	Example
Question	Does guidance change variety and subject clarity?
Keep fixed	Model, prompt, size, sampler, steps
Change	Guidance only
Repeat	Use the same set of seeds at each value
Judge	Subject clarity, appearance, and variety

L03 · 91 Lecture 03



L03 · 94



L03 · 62

Fix everything but one

Same model, prompt, size, sampler, steps; same seed list per guidance value.

Decide the test first

Judging questions before you look. Record failures; don't swap seeds.

Label what a figure shows

One run's progress ≠ finished runs at different step counts.

ASK YOURSELF

Is your A2 reconstruction log a fair comparison? Which variable leaked?

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

A Same prompt, a completely different composition every run

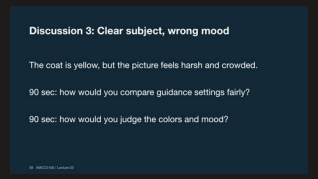
B Subject unmistakable, but colours burnt and every sample alike

C Looks fine, but 40 seconds per result, and you need it live

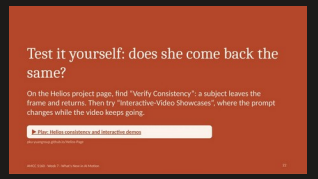
D A face changes after the character turns away and back

REVISIT

Click to open



L03 · 58



L07 · 22

Diagnosis: three knobs and one memory problem

A

Seed

Fix it to compare; record the whole workflow; use a reference for layout.

B

Guidance too high

Lower the scale; compare across the same seeds; one variable at a time.

C

Cost

Distilled model, fewer steps, quantization (Week 6). Measure seconds.

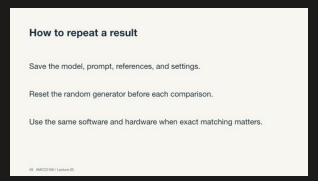
D

Not a knob: memory

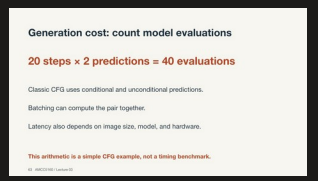
Representation and context over time. See Lectures 07 and 09.

REVISIT

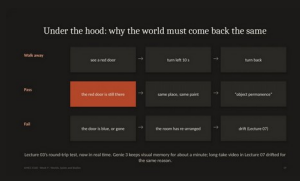
Click to open



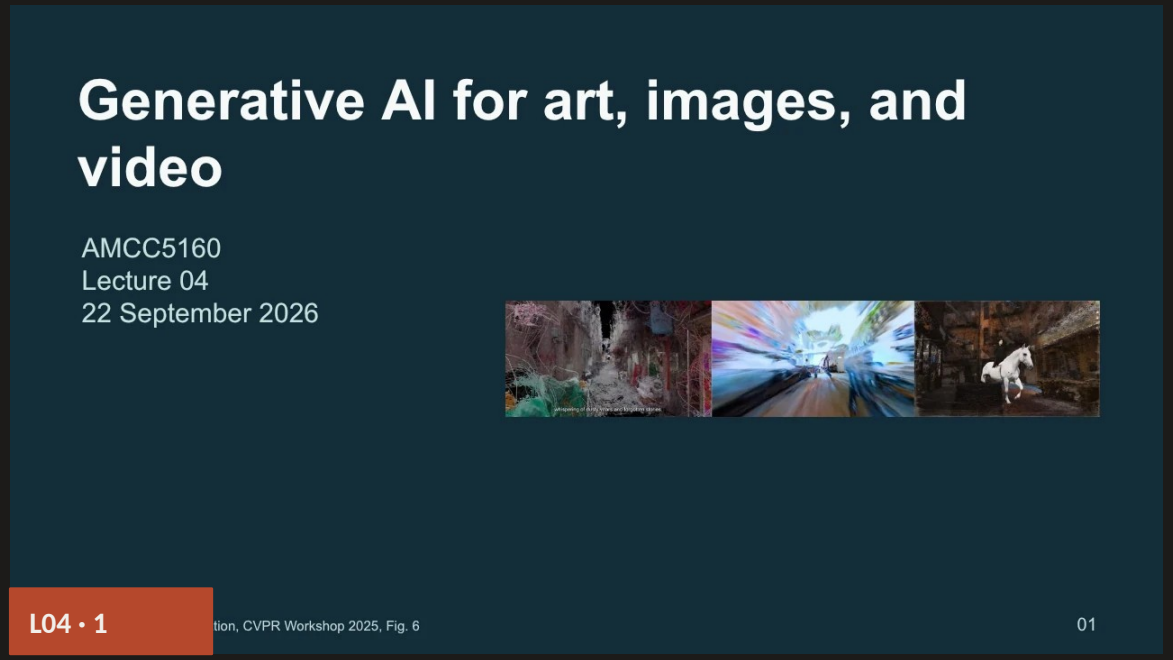
L03 · 49

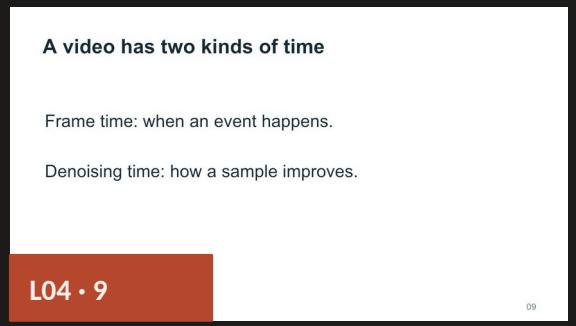


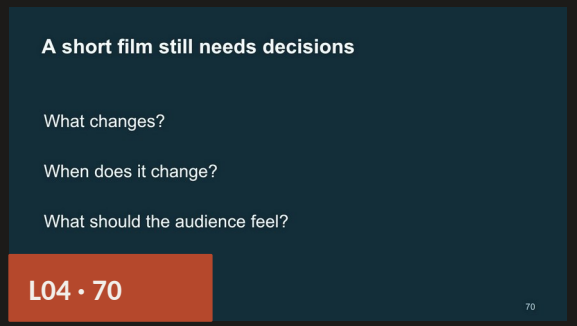
L03 · 63



L09 · 19







04

LECTURE 04 · 22 SEPTEMBER 2026

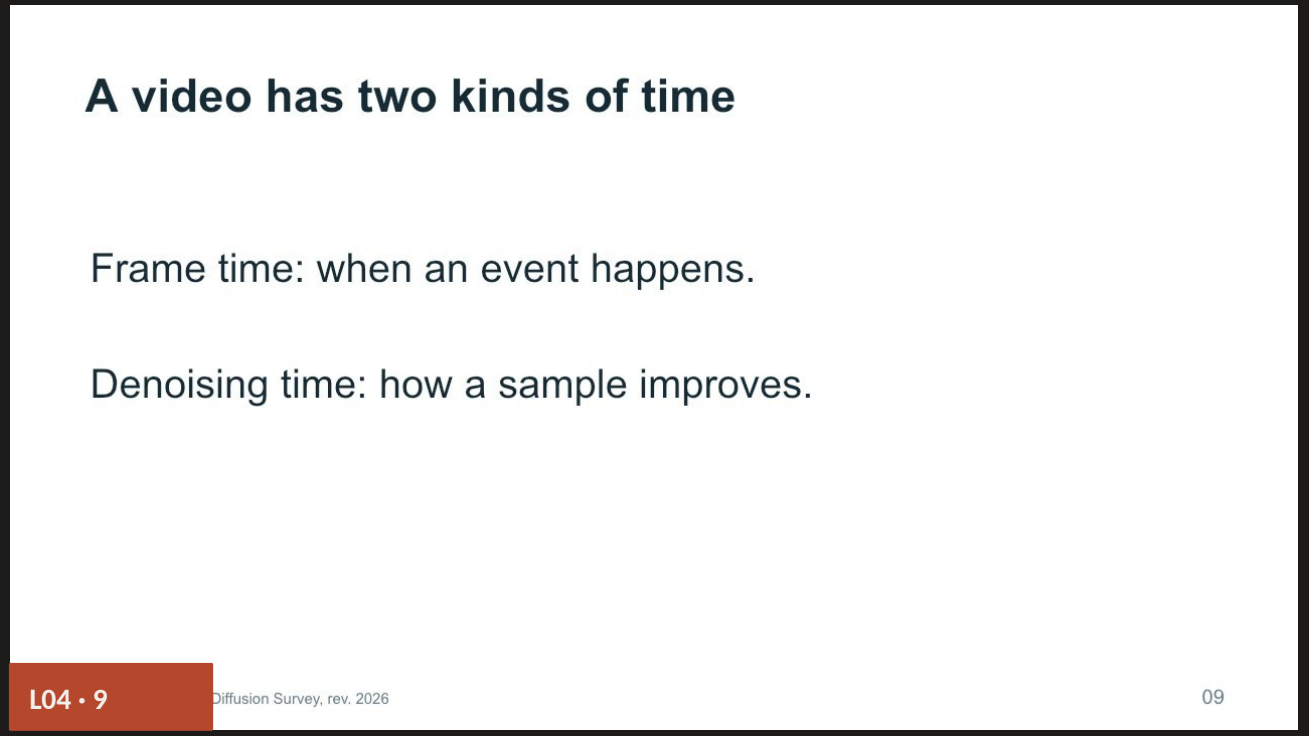
Video generation and art filmmaking

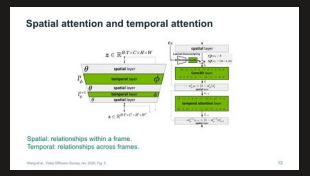
What changes when an image moves?

WE COVERED

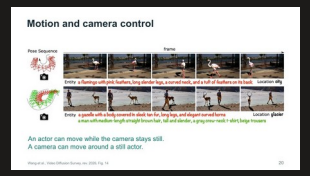
- Two kinds of time; spatial and temporal attention
- Training data, filtering and references
- Visual art: intention, scenario, modality
- Five art films and their workflows

Two kinds of time, two kinds of motion





L04 · 12



L04 · 20

Frame time ≠ denoising time

Scene time vs. computation steps. A step is not an extra frame.

Spatial vs. temporal attention

Within a frame vs. across frames. Connection is not understanding.

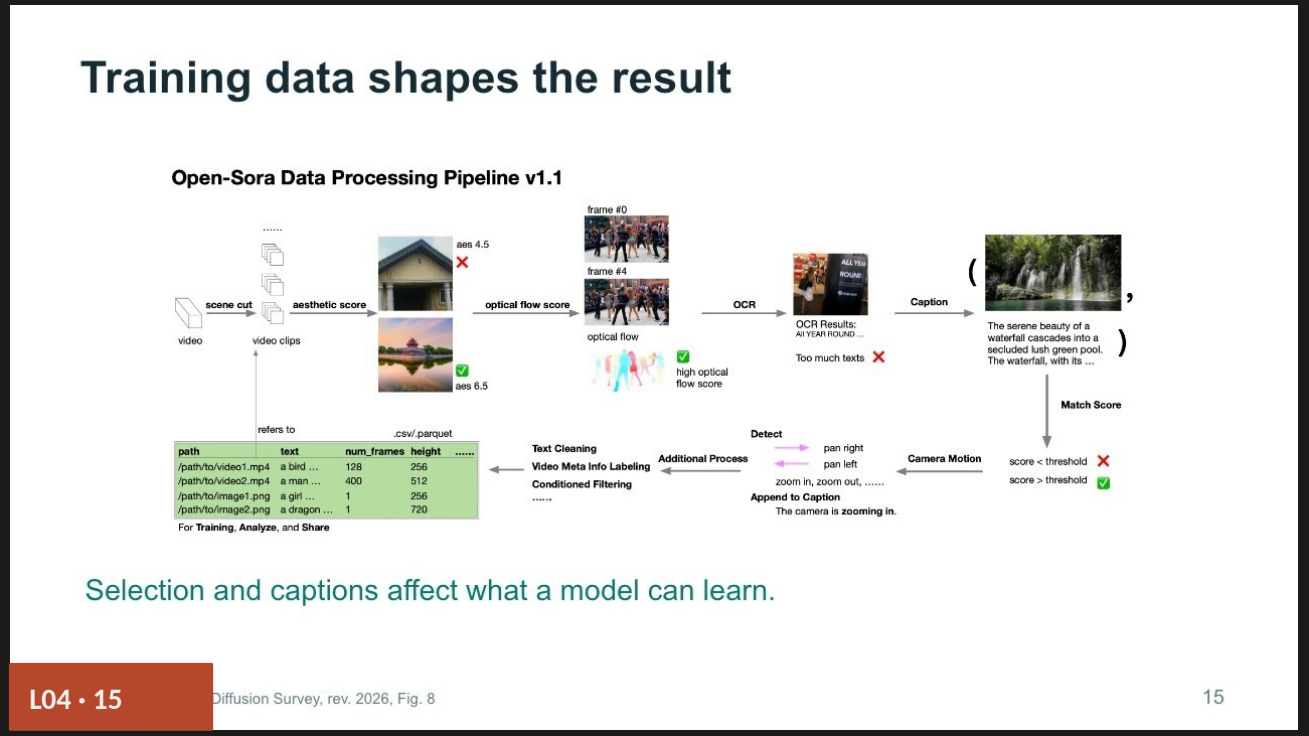
Object motion ≠ camera motion

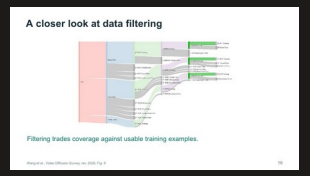
Both move pixels. Specify them separately to control them separately.

ASK YOURSELF

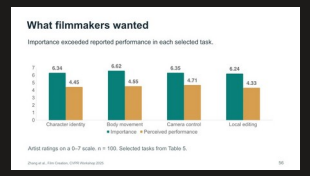
In your final project, which shot mixes object and camera motion?

Data shapes what a model can make





L04 · 16



L04 · 56

Every filter is a decision

Selection, filtering, captions: each stage changes what the model learns.

Gaps live in the data

If a visual language keeps failing, rewriting the prompt may not help.

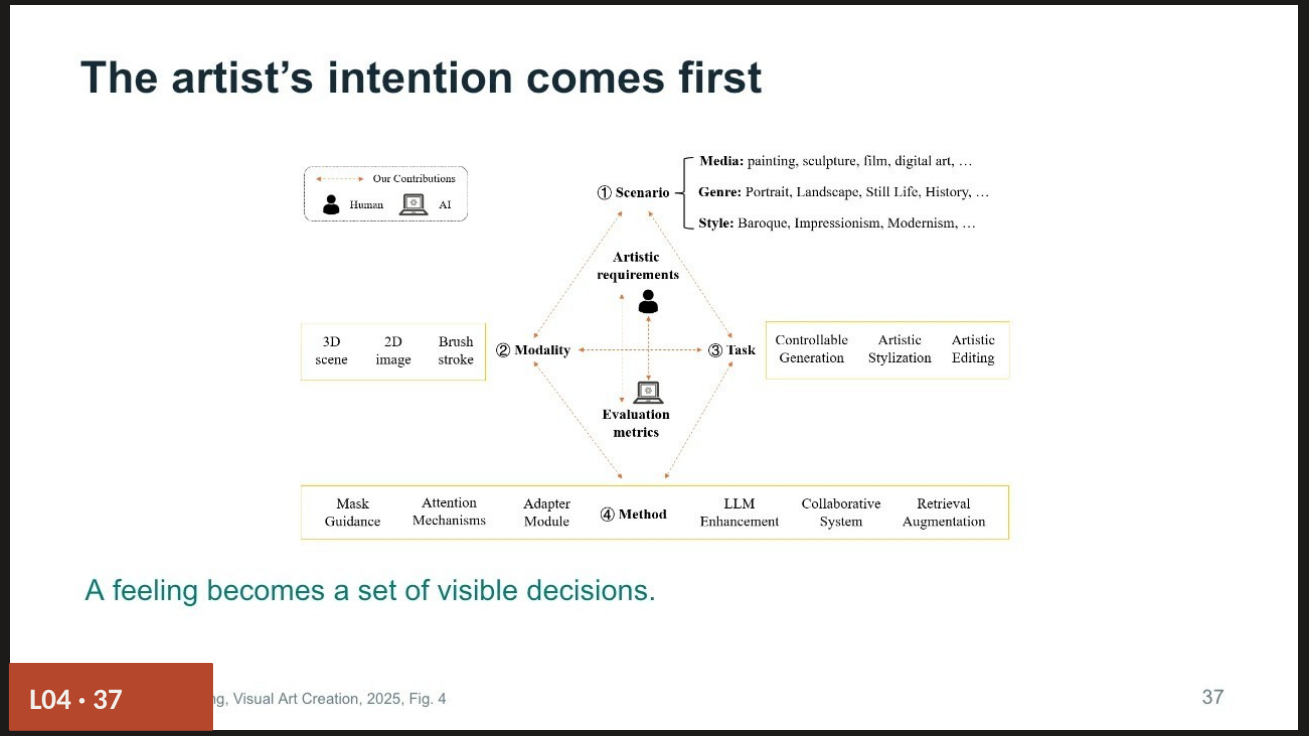
What filmmakers wanted

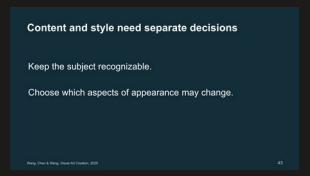
Identity, body movement, camera control, local editing: all showed a gap.

ASK YOURSELF

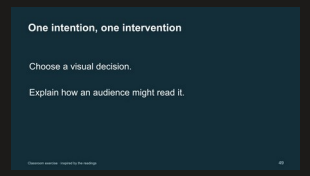
Where did a model's defaults resist your idea this term?

Intention first: one intention, one intervention





L04 · 43



L04 · 49

Connect intention to decisions

"Uneasy memory" → delayed movement, missing detail, one altered object.

Separate content and style

Keep silhouette, proportions, pose; let brushwork and palette vary.

One intention, one intervention

Then test it: show two versions and ask what people read.

ASK YOURSELF

State your final project's intention and its one visible intervention.

Five art films, five different workflows

Overthinking

Nostalgic look built across adaptation, assets, animation, edit

Clown

Inconsistency as expression: a fragmented identity

For Pixi

A recurring character via iteration and region editing

Fish Tank

Through 3D: UV maps and topology still done by hand

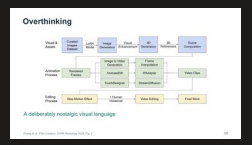
Volumography

Capture space; choose the camera afterwards

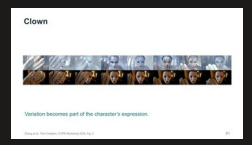
3.14 video-generation tools per film, on average, in the 2025 sample (L04 · 55)

REVISIT

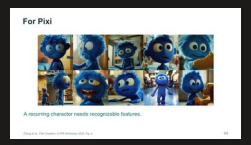
Click to open



L04 · 58



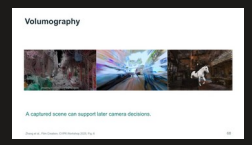
L04 · 61



L04 · 64

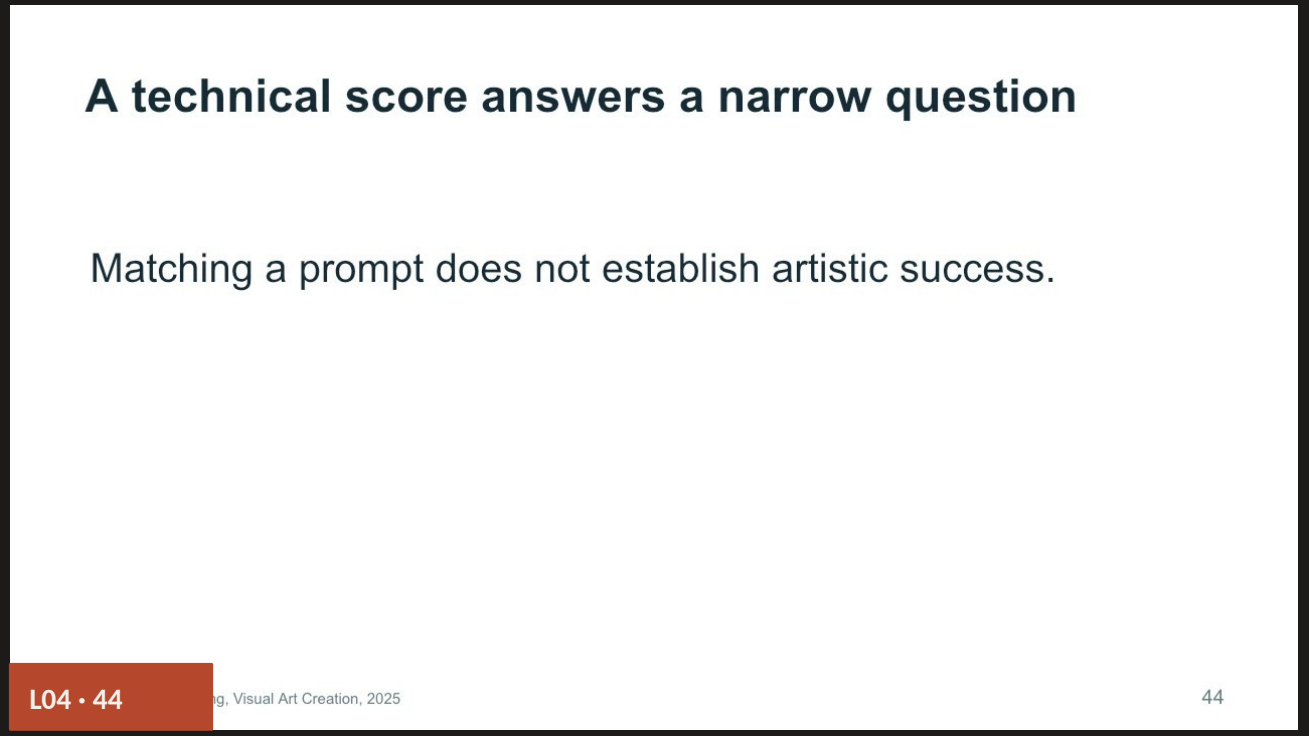


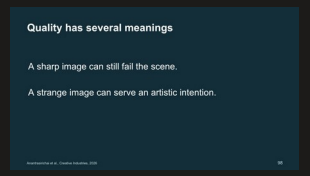
L04 · 66



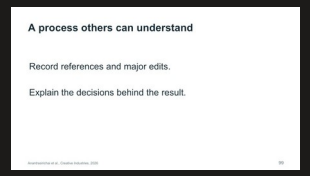
L04 · 68

Quality has several meanings





L04 · 98



L04 · 99

A score answers a narrow question

Alignment, similarity, perceptual quality: not whether the art works.

Judge separately

Technical continuity · clarity of the reveal · the audience's reading.

A process others can understand

Before/after versions, models, settings, and the decisions that changed it.

ASK YOURSELF

Name the three separate judgments you'll make about your final piece.

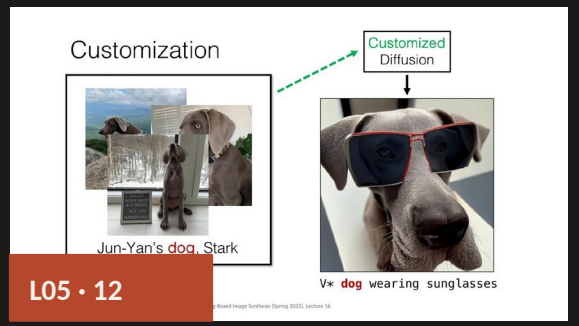
AMCC 5160 · LECTURE 05 · 29 SEPTEMBER 2026

Making the *Model Yours*

Adaptation, creative workflows, and authorship

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

L05 · 1



L05 · 12

BACK TO THE START

Whose model is it — now?

1. If a LoRA is only a few megabytes on top of a model trained on billions of images, how much of the output is yours?
2. A model can learn Greg Rutkowski's style and can be made to forget it. Who should decide which?
3. Would you share a LoRA of your own style? Under what terms?

L05 · 78

05

LECTURE 05 · 29 SEPTEMBER 2026

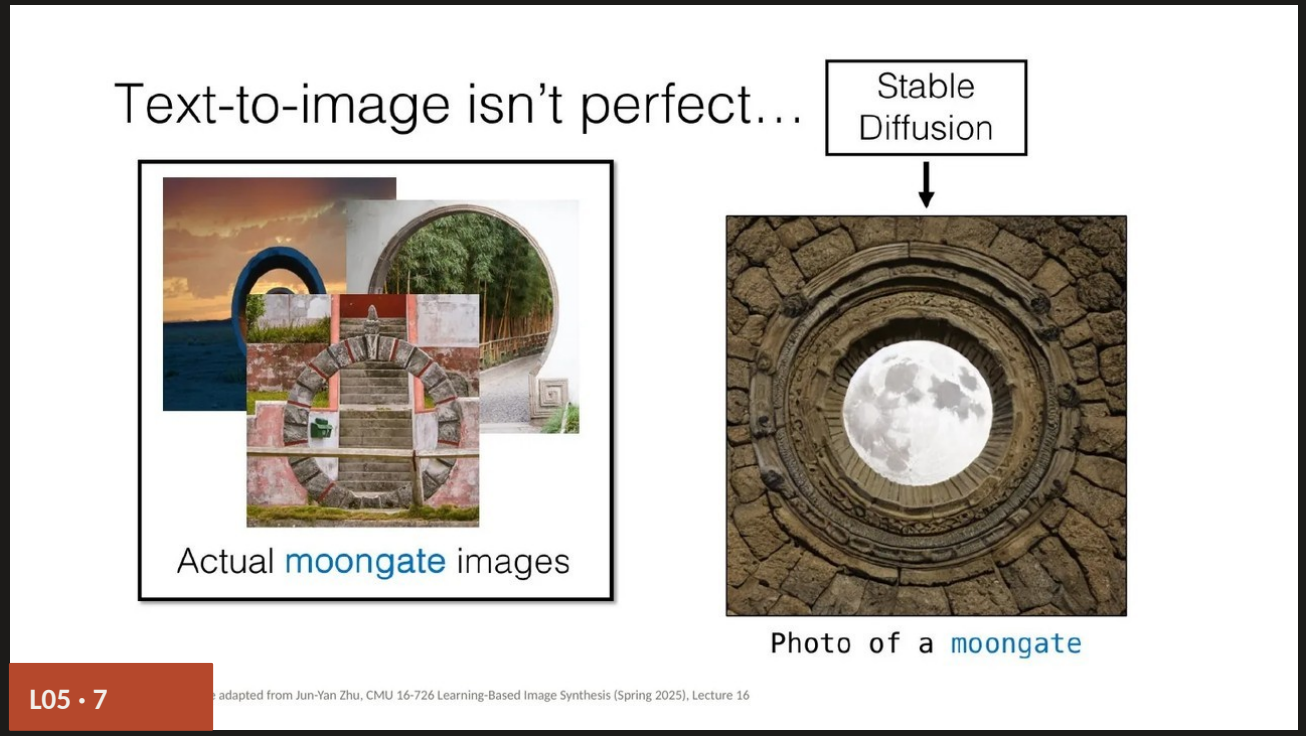
Making the model yours

Whose model is it?

WE COVERED

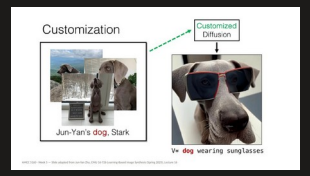
- What prompting alone cannot fix
- Textual Inversion, DreamBooth, Custom Diffusion, LoRA, IP-Adapter
- Memorization and concept ablation
- Copyright, provenance and bias

What prompting alone cannot fix

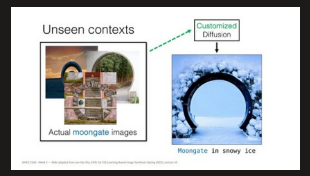




L05 · 11



L05 · 12



L05 · 10

Two bottlenecks

Language can't describe everything; data doesn't contain your things.

A description is not an identity

“Dark grey weimaraner” gives a weimaraner, never Stark. That's A1's drift.

A new word: V*

Teach one token, then recombine: a moongate in snow, Stark in sunglasses.

ASK YOURSELF

What in your project could no prompt describe?

Making the model yours: what actually gets trained?

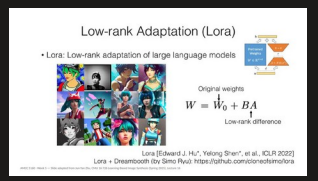
Method	What is trained	Best for · cost
Textual Inversion	One new word embedding; model frozen	Styles · tiny file, weak identity
DreamBooth	All weights, with class-prior regularization	Identity · heavy, can overfit
Custom Diffusion	Only K and V in cross-attention	Combining concepts · small update
LoRA	A low-rank overlay: 16.8 M → 131 k per layer	The artist's default · swap overlays
IP-Adapter	Nothing per subject: an image encoder	Instant references · less faithful

REVISIT

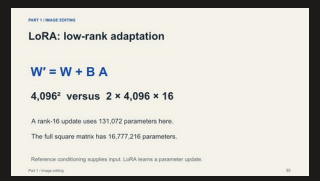
Click to open



L05 · 27



L05 · 61



L02 · 35

Overfit, memorize, forget

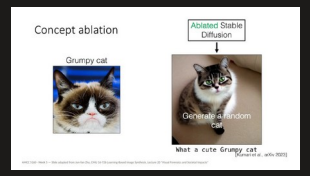




L05 · 28



L05 · 54



L05 · 45

Overfit

Moongates in every moon. Fix: regularization images of what to keep.

Memorize

Duplicated training images come back almost exactly; styles too.

Forget

Concept ablation: Grumpy Cat → a cat; Rutkowski's look removed, boat kept.

ASK YOURSELF

Does your training or reference set contain duplicates?

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

Full update

$$W' = W + \Delta W$$

$4096 \times 4096 = 16{,}777{,}216$

Every entry of ΔW is learned

LoRA, rank r = 16

$$W' = W + B \cdot A$$

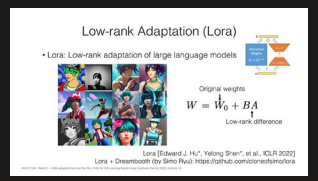
$2 \times 4096 \times 16 = 131{,}072$ (128× fewer)

W stays frozen; swap overlays, keep the model

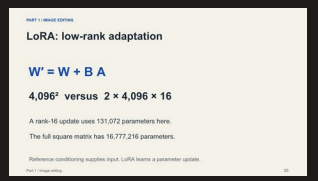
Small update ≠ general concept. Test on poses and settings the training images never showed.

REVISIT

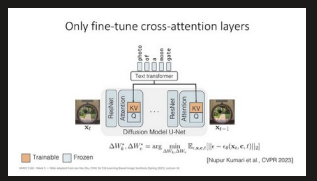
Click to open



L05 · 61

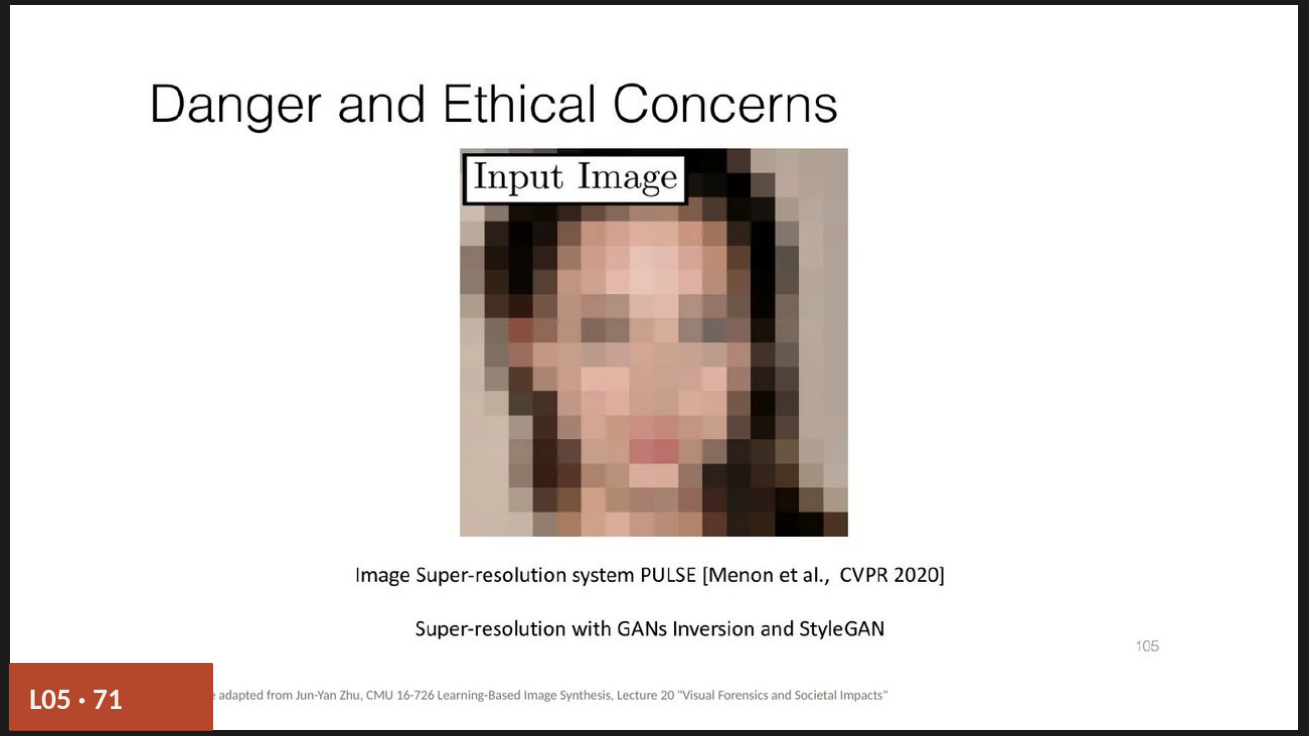


L02 · 35

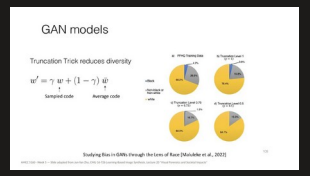


L05 · 34

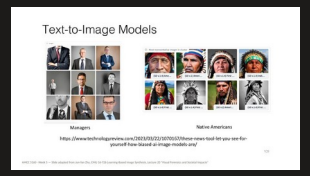
Defaults have politics



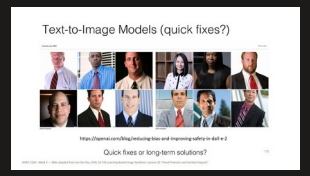
L05 · 71



L05 · 74



L05 · 75



L05 · 76

Details come from the prior

PULSE invents a sharp face; the missing details come from the training data.

A quality knob with a social effect

Truncation pulls toward the average and shrinks diversity.

Quick fixes or your own data?

Prompt rewriting helps; building your own dataset goes further.

ASK YOURSELF

Whose culture is missing or flattened in the model you used most?

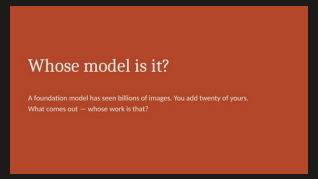
“Whose model is it?”

The company that trained it, you who adapted it, or the people whose images are inside it?

Show of hands, 29 September. Has your hand moved?

REVISIT

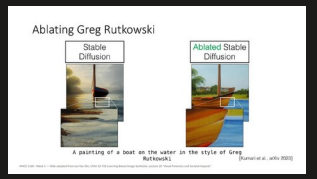
Click to open



L05 · 3



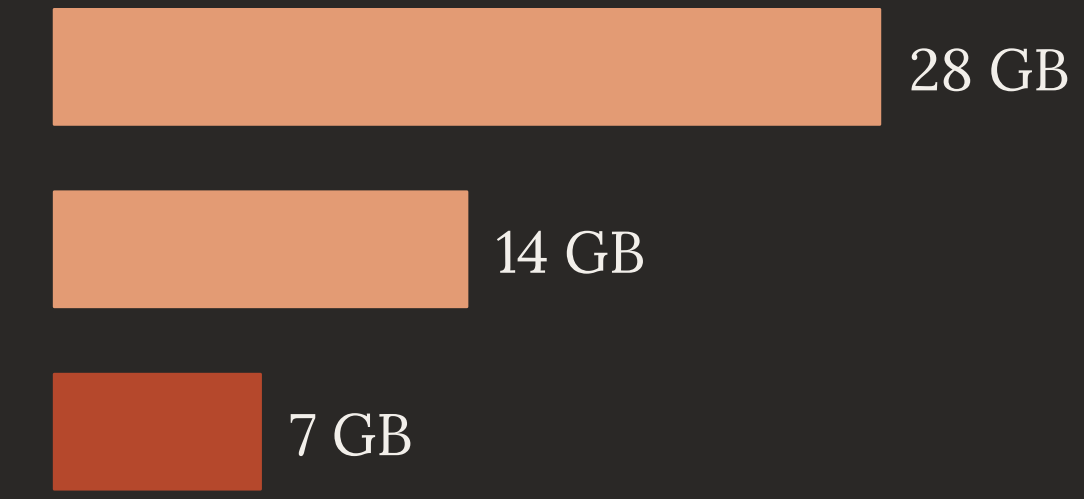
L05 · 30



L05 · 54

Week 6: store less, so you can run it

MEMORY FOR THE WEIGHTS OF A 14 B MODEL



14 B parameters, weights only (L07 · 61). Activations, VAE and text encoder need memory on top.

Change less

L05

Adapters · LoRA

Step less

L07

Distillation

Store less

L06

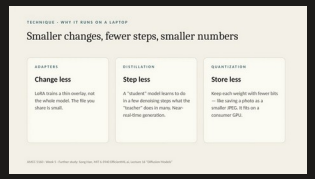
Quantization

REVISIT

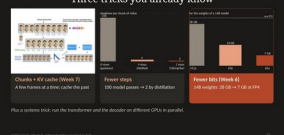
Click to open



L07 · 61



L05 · 67



L08 · 65

AMCC 5160 · LECTURE 07 · 13 OCTOBER 2026

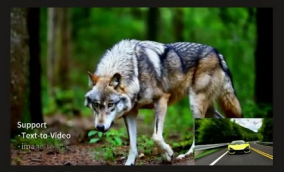
What's New in *AI Motion*

Music, minutes, and machines · Fall 2026

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

L07 · 1

Helios: a minute of video, in real time



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

16

L07 · 16

Eighteen months as a product

Dec 2024	Sep 30, 2025	Mar 24, 2026	Apr 26, 2026	Sep 24, 2026
Sora launches as a product	Sora 2 + a social video app	OpenAI announces the shutdown	App and website close	API switched off, no replacement named

OpenAI: What to know about the Sora discontinuation ↗

59

L07 · 59

07

LECTURE 07 · 13 OCTOBER 2026

Music, minutes, and machines

What happens when generation gets long, fast and physical?

WE COVERED

- 30-second scenes and real-time minutes (Helios)
- Wan-Dancer: song in, dance out
- Phone video teaches a humanoid to dance
- Sora's shutdown: keep your sources

From the shot to the scene

~5 s

2024

A typical clip

~15 s

Early 2026

Seedance 2.0

30 s

Summer 2026

Seedance 2.5 · Wan 3.0 · Kling 4.0

~1 min

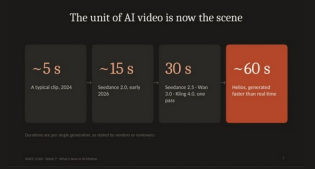
Real time

Helios: ~19.5 fps on one H100

All at once: consistent, expensive. Forward in time: streams, but must fight drift.

REVISIT

Click to open



L07 · 7

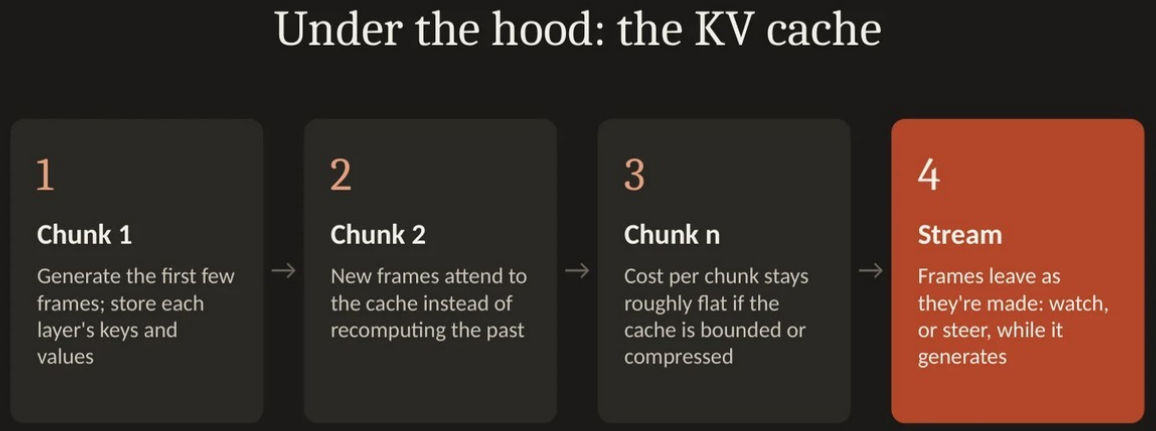


L07 · 10

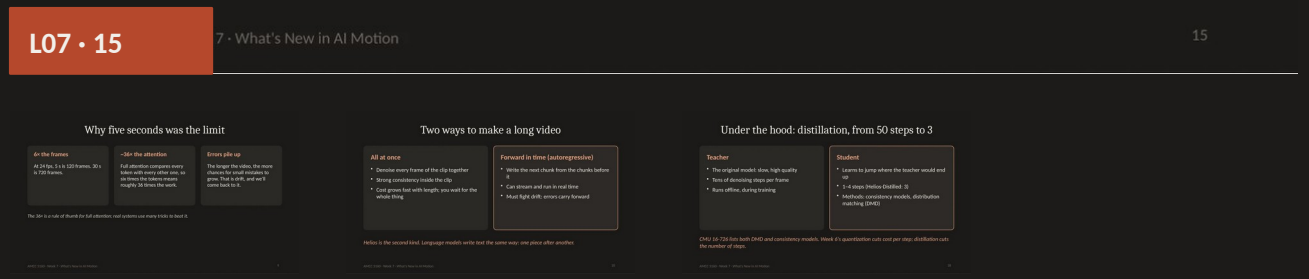


L07 · 16

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



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



L07 · 9 L07 · 10 L07 · 18

Why five seconds was the wall

6× the frames → about 36× the attention comparisons.

Write forward + KV cache

Chunks stream like words; the cached past is not recomputed.

Distillation

A student jumps where the teacher lands: Helios ~50 → 3 steps.

ASK YOURSELF

Which of these three would make your project faster, and what would it cost?

Music becomes the choreographer

The trick: plan the song, then the steps

Global level: the whole song

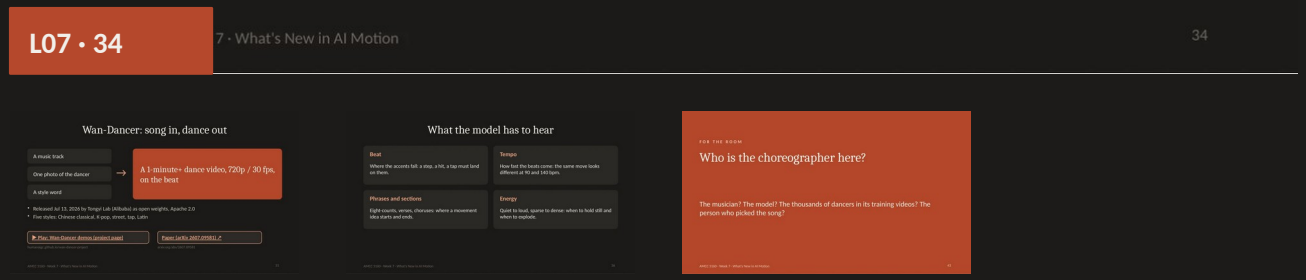
Intro	Verse	Build	Chorus	Outro

↓ plans

Local level: the moves

Renders each passage's steps, beat by beat

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



L07 · 31 L07 · 36 L07 · 43

Song + photo + style → a minute

Wan-Dancer: open weights, five styles, on the beat.

Plan the song, then the steps

Global structure first, local steps second, like a choreographer.

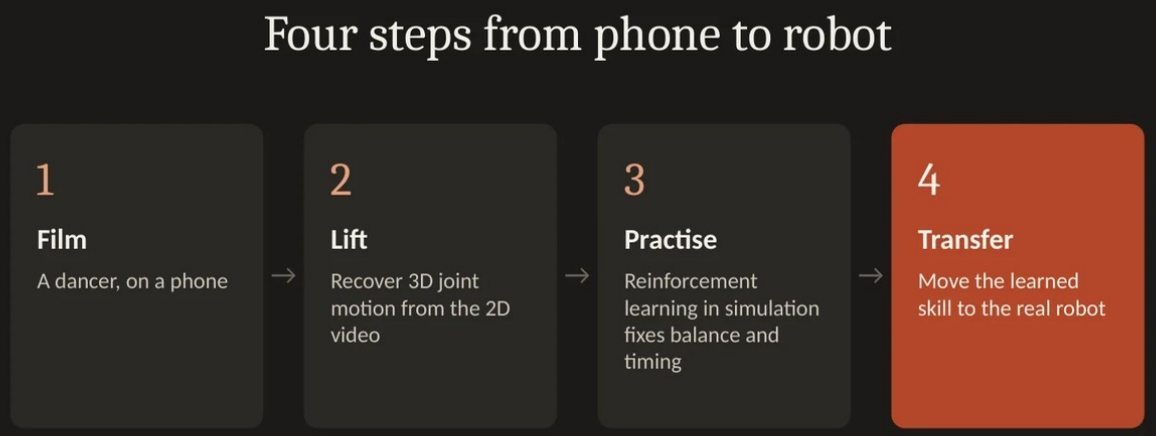
Limits

One photo invents the back; whose dances were in the training data?

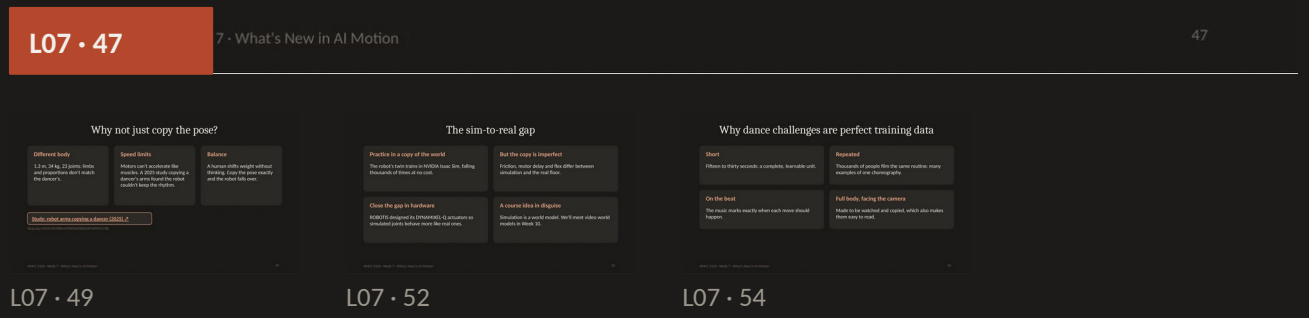
ASK YOURSELF

Who is the choreographer here?

Motion leaves the screen



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



L07 · 49 L07 · 52 L07 · 54

Phone → 3D → practise → transfer

Copying the pose fails: different body, slower motors. Retarget, then RL.

The sim-to-real gap

Friction, delay, flex never match. Domain randomization helps.

Video trains bodies

Dance challenges are near-perfect data. Your post has two audiences.

ASK YOURSELF

If a generated video teaches a real body to move, is it still “fake”?

Tools vanish. Keep your sources.

```
my-project/
  README.md      model · version · date · settings · one-line intent
  prompts/       every prompt kept, word for word, with its seed
  refs/          images · clips · songs, with credits and licences
  raw/           unedited generations, interesting failures too
  edit/          the edit project and the exported master
```

Dec 2024

Sora becomes a product

24 Sep 2026

API switched off, no replacement

~18 months

as a product

REVISIT

Click to open



L07 · 59



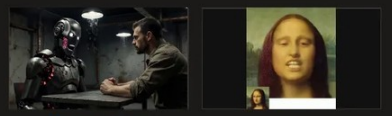
L07 · 65

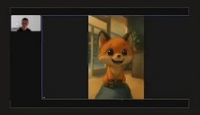


L07 · 66

AMCC 5160 · LECTURE 08 · 20 OCTOBER 2026

When AI Video *Learned to Talk*



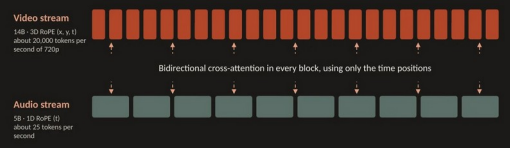



Sound, performance, and real-time characters · Fall 2026

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

L08 · 1

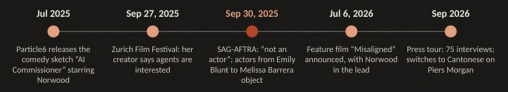
Inside LTX-2: two streams, one clock



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 [illegible]233.

L08 · 19

Tilly Norwood: an “actress” who never existed



L08 · 85

08

LECTURE 08 · 20 OCTOBER 2026

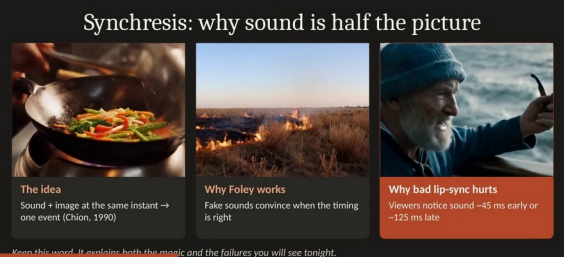
When AI video learned to talk

Who directs a performance when one model makes the sound and the picture?

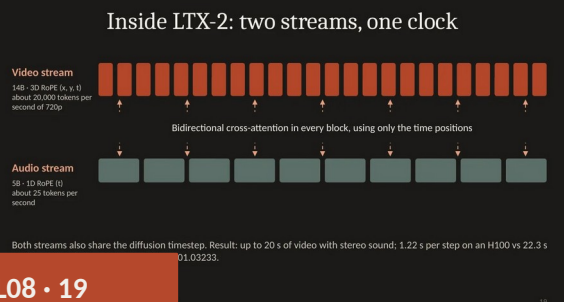
WE COVERED

- Native audio: two streams, one clock
- Lip-sync, visemes and the scoreboard trap
- Real-time characters that listen
- Synthetic performers, consent and the law

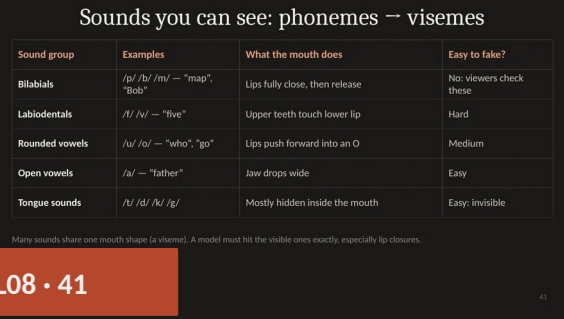
Sound shares the clock



L08 · 6



L08 · 19



L08 · 41

1 **Synchresis (Chion, 1990)**

Sound and image at the same instant weld into one event. Off-sync is felt by everyone. (L08 · 6)

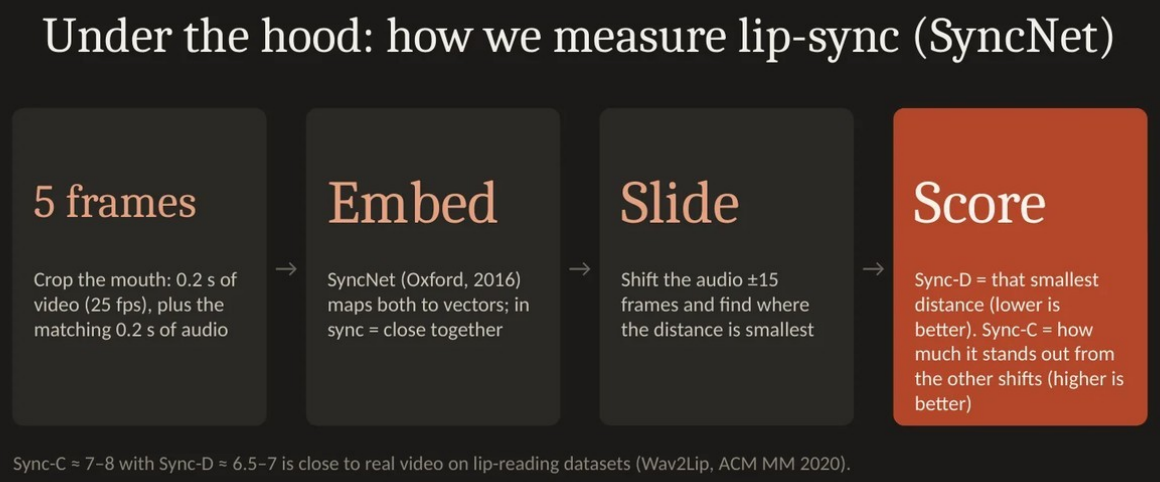
2 **Two streams, one clock**

LTX-2: 14 B video + 5 B audio; cross-attention on time positions only. (L08 · 18–21)

3 **Check p, b and m**

Bilabials need closed lips. The fastest lip-sync test there is. (L08 · 41)

Reading a lip-sync score, and its trap



L08 · 43 8 · When AI Video Learned to Talk 43



L08 · 41



L08 · 44

Eye test: p, b, m

Bilabials need closed lips. Tongue sounds are hidden; open vowels are easy.

SyncNet: Sync-D and Sync-C

Slide audio ±15 frames; lowest distance, and how much it stands out.

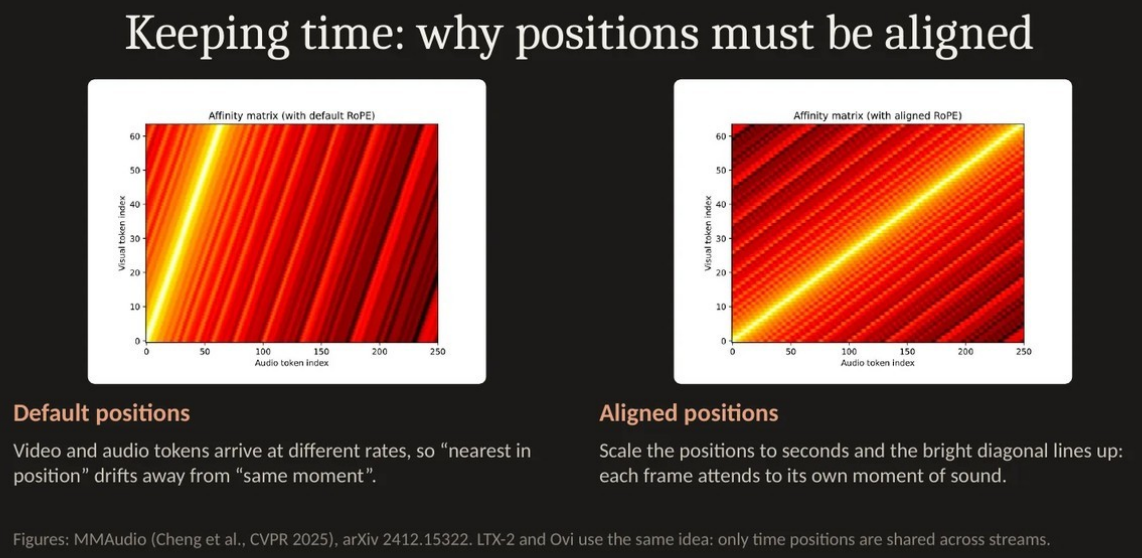
The trap

Wav2Lip 6.4 vs real 6.7 Sync-D: trained on the judge, so it “beats” reality.

ASK YOURSELF

Which number in your project could a model have been trained to please?

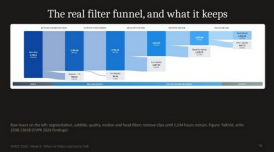
Sync starts in the data and in the positions



L08 · 21



L08 · 26



L08 · 78

Align positions to real time

Naive numbering misaligns streams; scaled or temporal RoPE gives a clean diagonal.

Filter for real sync

Drop pasted music, dubbing, off-screen voices; caption the sound, quote the words.

TalkVid: 6,000+ → 1,244 hours

Speech, quality, motion, head filters; ~95% agreement with human judges.

ASK YOURSELF

Write your next prompt with the sound described and the spoken words in quotes.

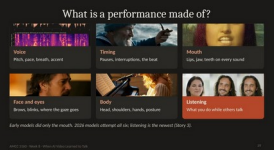
Directing a performance, not a prompt

Directing the delivery: what to ask for

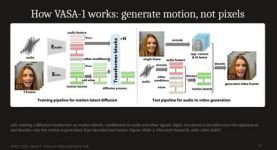
Direction	Voice	Face and body	Prompt words that help
Tired	Slower, lower, breathy	Heavy lids, slow blinks, little head motion	“quietly, exhausted, after a long pause”
Angry, held in	Clipped, even pitch	Tight jaw, fixed stare	“through clenched teeth, without raising his voice”
Nervous	Fast, rising ends, 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.

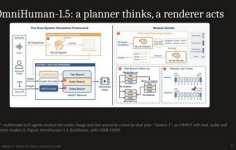
L08 · 57 8 · When AI Video Learned to Talk 57



L08 · 39



L08 · 47



L08 · 51

Six parts of a performance

Voice, timing, mouth, face and eyes, body, and listening.

Generate motion; plan intention

VASA-1 renders a motion code; OmniHuman-1.5 plans, then acts.

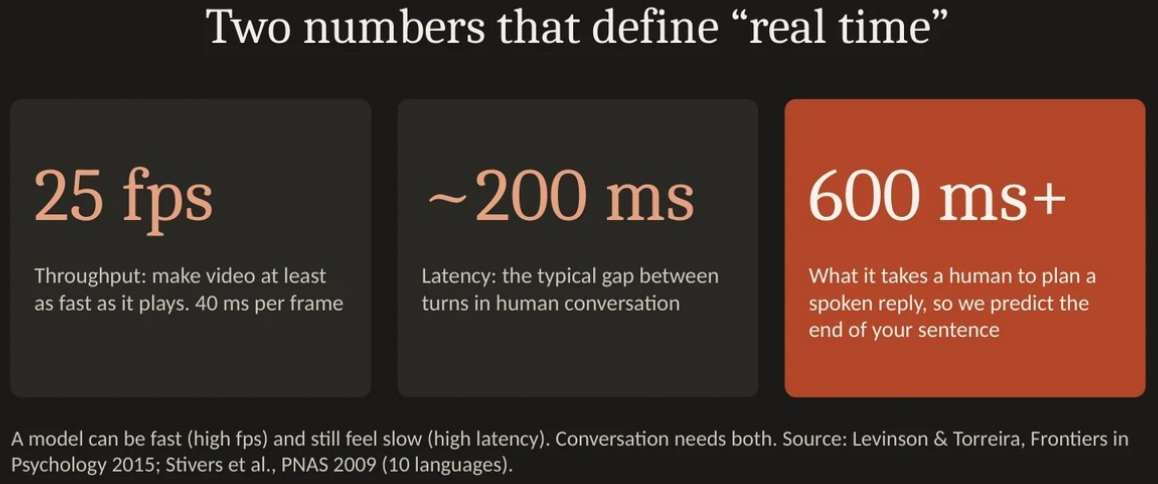
Playable actions, not labels

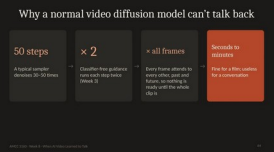
“Through clenched teeth, without raising his voice.” Or record the line yourself.

ASK YOURSELF

Rewrite one emotion label in your project as a playable action.

Characters that listen: the real-time budget





L08 · 64



L08 · 65



L08 · 71

Two numbers

Throughput: 40 ms per frame. Latency: humans leave ~200 ms between turns.

Why diffusion can't talk back

30–50 steps × 2 for guidance × bidirectional attention.

Fixes you already knew

Chunks + KV cache, distillation (24 → 2), quantization. Then: listen.

ASK YOURSELF

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

The real-time toolkit, gathered from four lectures

Change less

Adapters · LoRA

A thin overlay instead of the whole model

L05

Step less

Distillation

Helios ~50 → 3 steps · TalkingMachines 24 → 2 evaluations

L07 · L08

Store less

Quantization

8 or 4 bits per number

L06 (TA talk)

Remember smartly

Chunks + KV cache

Look backwards only; don't recompute the past

L07 · L08

40 ms

per frame at 25 fps

~200 ms

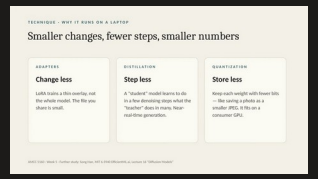
human turn gap

0.07 → 100+

fps: NeRF → splats

REVISIT

Click to open



L05 · 67



L07 · 15



L08 · 65

Who is the performer?



A performer who never existed

Tilly Norwood: backlash in 2025, a lead role announced in 2026.

SAG-AFTRA 2026

Synthetic performers: a strong disincentive, not a ban. Replicas: consent and pay.

Same tools, opposite receptions

A non-consensual viral fight vs. a disclosed Tribeca premiere.

L08 · 85 8 · When AI Video Learned to Talk 85



L08 · 88



L08 · 90

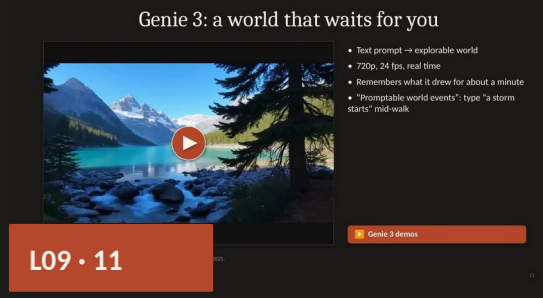


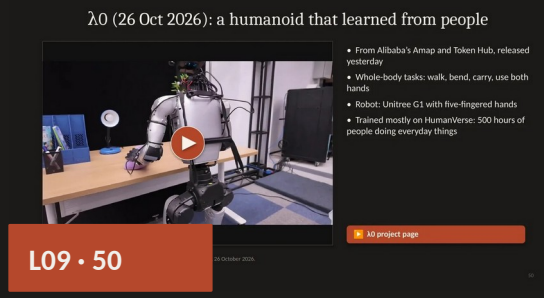
L08 · 92

ASK YOURSELF

An AI performer that never existed should be allowed to win an acting award. Yes or no?







09

LECTURE 09 · 27 OCTOBER 2026

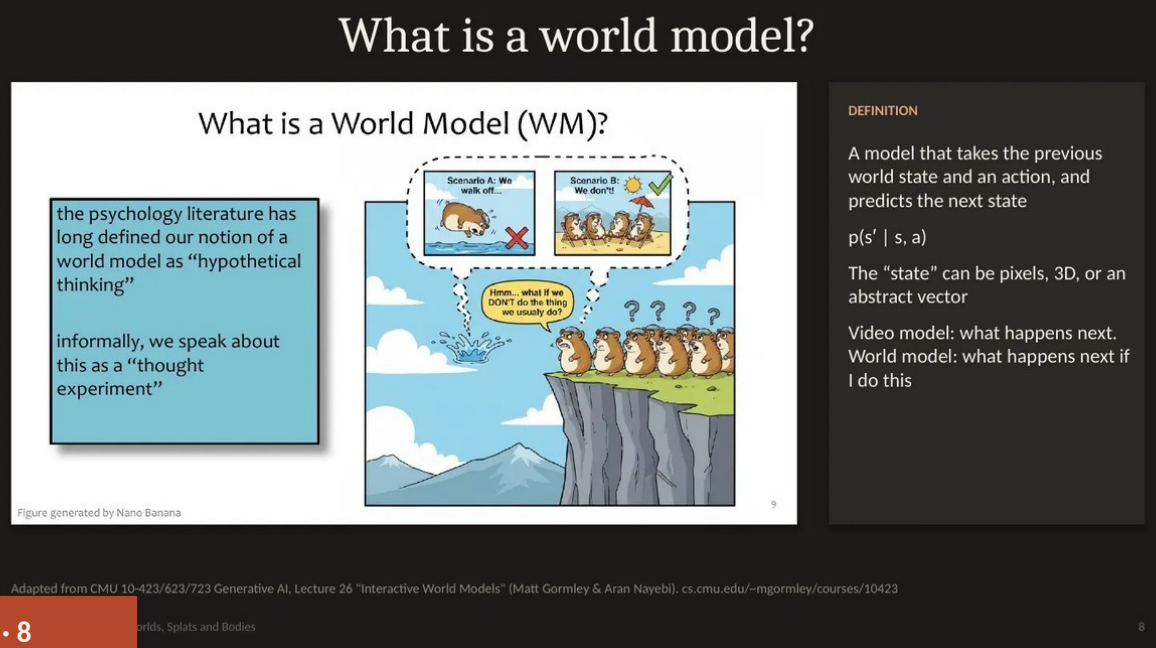
Worlds you can walk into, robots that learn by watching

What happens next if I do this?

WE COVERED

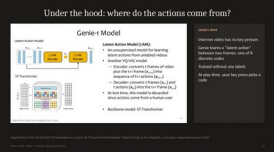
- Genie 3: a world that waits for you
- Gaussian splats: from NeRF to a file format
- VLAs, λ0 and learning from human video
- Pixels or meaning: Cosmos vs. JEPA

Add an action, and video becomes a place





L09 · 15



L09 · 16



L09 · 17

State + action → next state

One extra input turns film into game; the model must wait for you.

Latent actions

Squeeze the change between two frames into one of 8 codes: left, right, jump.

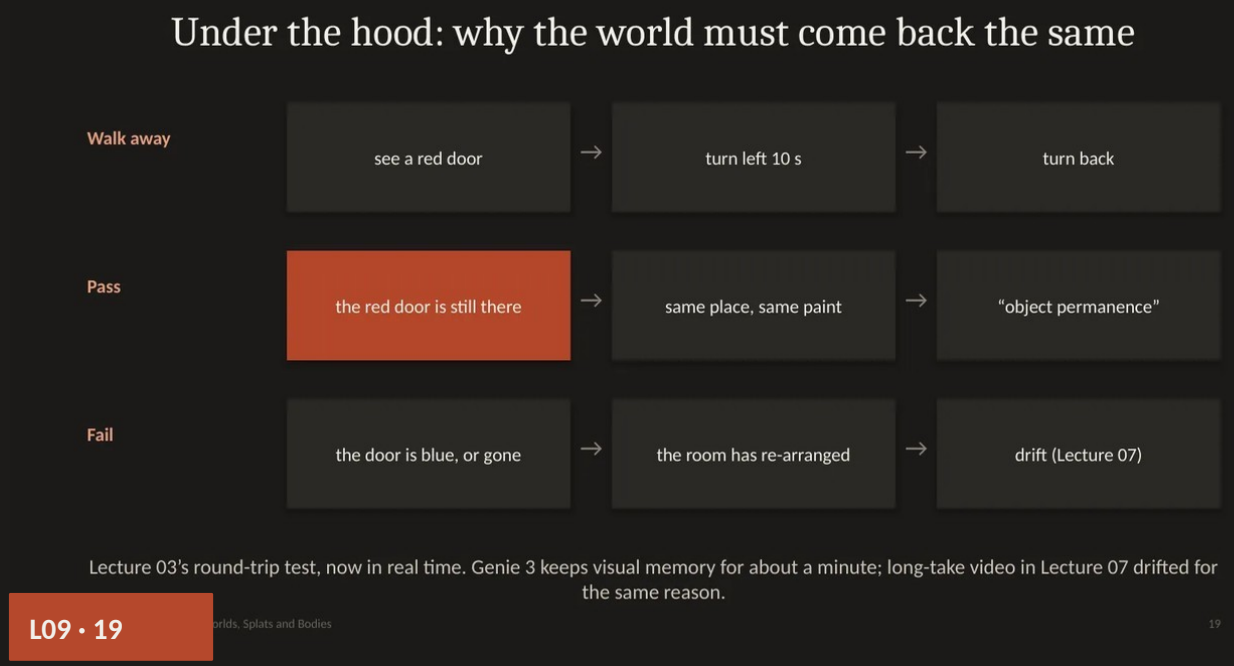
A loop at 24 fps

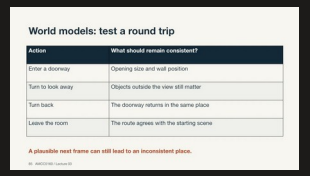
Tokenize → action → predict → decode. Genie 3: 720p, real time, world events.

ASK YOURSELF

If a world only exists while someone plays it, what is the artwork?

Memory: walk away and come back





L03 · 85



L09 · 13

The red-door test

Turn away, turn back. Weak models forget what left the context window.

Test what you claim

Round trips check spatial agreement, not contact, materials or control.

Serious uses

Waymo's Genie 3 variant generates camera and lidar for rare situations.

ASK YOURSELF

Run the round-trip test on any world or game in your A3.

Drift: one problem in three costumes

Long video

Lecture 07

Trained on perfect frames, generating from its own. A jacket's shade drifts by second 60.

Self Forcing · Helios trains on drifted video

World models

Lecture 09

Turn away from the red door, turn back: it is blue. Memory is the context window.

Genie 3: about a minute of memory

Robots

Lecture 09

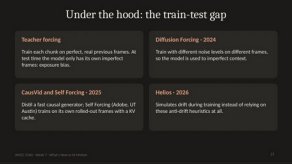
A small error leads somewhere the expert never went. Errors compound.

Practice (RL) on top of copying

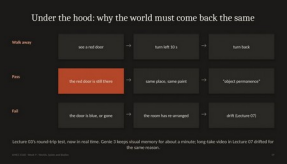
Train on clean data, then live on your own mistakes. The fix: train on your own outputs, remember, practise.

REVISIT

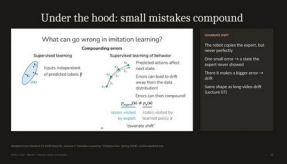
Click to open



L07 · 21



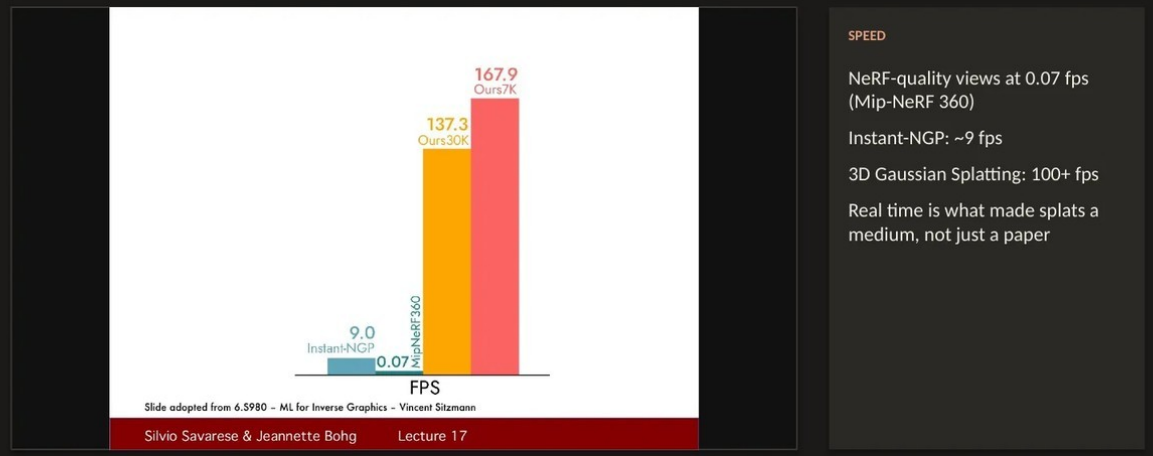
L09 · 19



L09 · 48

3D is back: from NeRF to Gaussian splats

The result: from 0.07 to 100+ frames per second



SPEED

NeRF-quality views at 0.07 fps (Mip-NeRF 360)

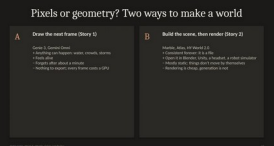
Instant-NGP: ~9 fps

3D Gaussian Splatting: 100+ fps

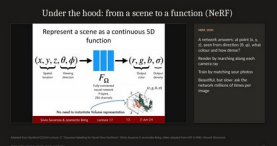
Real time is what made splats a medium, not just a paper

Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

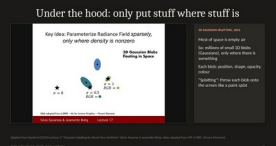
L09 · 37



L09 · 27



L09 · 33



L09 · 34

Pixels or geometry

Draw the next frame (alive, forgetful) or build a file (consistent, static).

NeRF → splats

Ask a network everywhere, or put blobs only where stuff is.

0.07 → 9 → 100+ fps

Now a normal format: Blender 5.3, three.js, a glTF extension.

ASK YOURSELF

A photo freezes a moment; a splat freezes a place. What should you be allowed to scan?

Generate the small thing, render the big thing

VASA-1

Lecture 08

Generate a tiny motion code for pose and expression; render the face from one photo.

40 fps

Genie latent actions

Lecture 09

Squeeze the change between two frames into one of 8 codes. Nobody labels anything.

8 codes → left, right, jump

Gaussian splats

Lecture 09

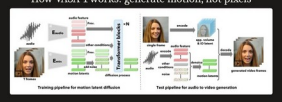
Put soft coloured blobs only where something exists; skip the empty air.

0.07 → 100+ fps

When your project is too slow or unstable, ask: what is the small thing?

REVISIT

Click to open



L08 · 47



L09 · 16



L09 · 34

Robots that learn by watching people

λ0, by the numbers

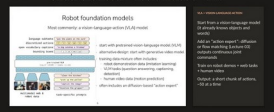
40.0%	62.5%	22.5%	15.0%
success of the strongest baseline (π0.5) on 4 real tasks	λ0: 25 of 40 real trials	λ0 without stage II (no human-to-robot translation)	λ0 without any human-data pretraining

The ablation is the lesson: remove the human video and success falls from 62.5% to 15%. Human video is not decoration; it is most of the skill.

L09 · 53

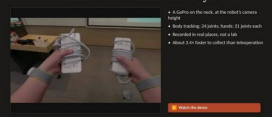
orlds, Splats and Bodies 53

Under the hood: what is inside a VLA?



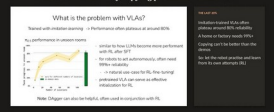
L09 · 46

HumanVerse: wear the robot's eyes



L09 · 51

Under the hood: why copying plateaus around 80%



L09 · 59

A VLA = VLM + action expert

Flow or diffusion outputs a chunk of joint commands, not pixels.

Wear the robot's eyes

HumanVerse: head-height camera, body and hand trackers, everyday tasks.

62.5% vs 40%

Without human data: 15%. Copying plateaus near 80%; homes need 99%.

ASK YOURSELF

Whose bodies and homes become the training data?

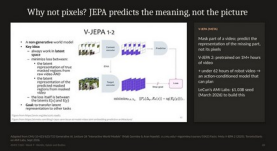
Does a world model need to draw pixels?



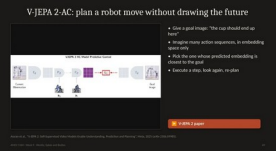
L09 · 65



L09 · 66



L09 · 68



L09 · 69

Generate the world

Genie, Omni, Atlas, Cosmos 3: watchable, trainable inside, sellable as video.

Predict the meaning

JEPA: predict the summary of what's missing, not its pixels.

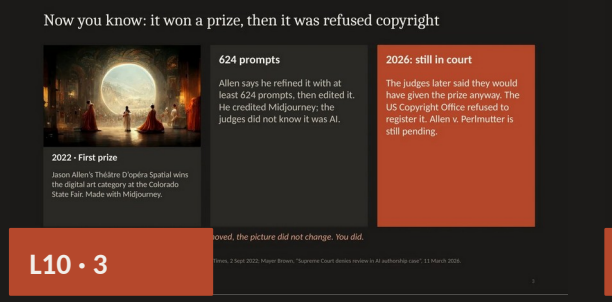
Plan without drawing

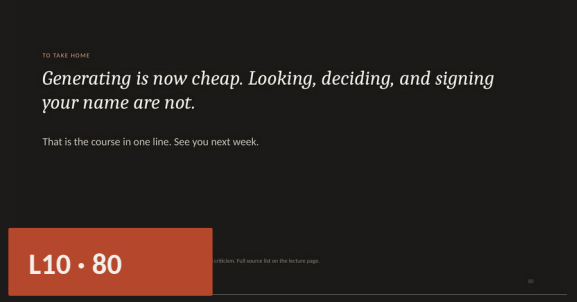
V-JEPA 2-AC picks actions in embedding space, steps, re-plans.

ASK YOURSELF

A video generator is a world model. For or against?







10

LECTURE 10 · 3 NOVEMBER 2026

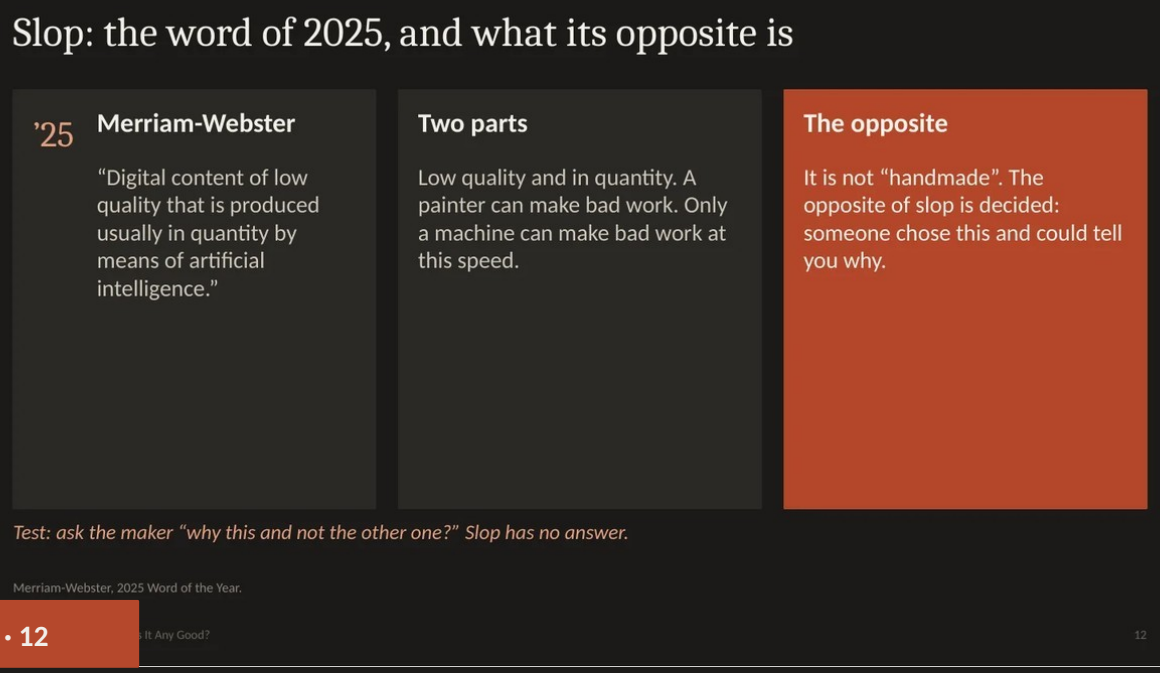
Is it any good?

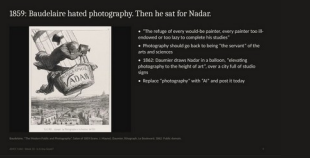
Is it any good, and how would you know?

WE COVERED

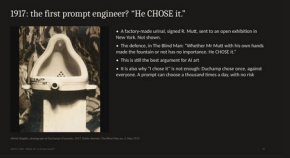
- Looking: critique methods and reading the dataset
- Measuring: PSNR, FID, VBench, Elo, aesthetic scores
- Showing: Dataland, Madison, the V&A
- Signing: C2PA, watermarks, the credit block

Every new medium gets asked: is this even art?





L10 · 9



L10 · 10



L10 · 11

An old question

Baudelaire 1859 · Duchamp 1917 · Molnár 1969 · Pope Leo, October 2026.

"He CHOSE it."

The best argument for AI art, as long as the choice costs you something.

The opposite of slop

Not handmade: decided. Someone chose, and can say why.

ASK YOURSELF

Where in your project can you show that something was decided?

Four questions people mix up

1

Is it AI?

Fact · evidence

Credentials, watermarks, logs

L10

2

Is it good?

Judgment · looking

Critique; scores are narrow

L02 · L04 · L10

3

Is it fair?

Ethics · consent

Data, likeness, places, jobs

L05 · L08 · L09

4

Is it allowed?

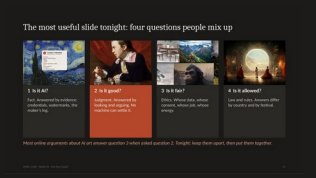
Law · rules

Lawsuits, SAG-AFTRA, EU Art. 50

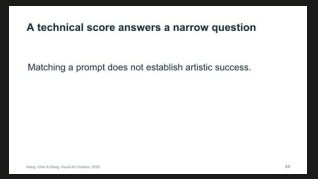
L05 · L08 · L10

REVISIT

Click to open



L10 · 13



L04 · 44

Look before you judge

See · Think · Wonder

Harvard Project Zero

1. What do you see? Facts only
2. What do you think about it?
3. What does it make you wonder?

Feldman's four steps

Classic art criticism

1. Describe
2. Analyze
3. Interpret
4. Judge, at the very end

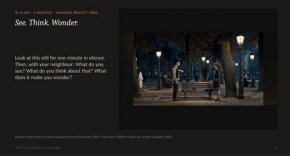
Reading an AI image

Salvaggio, Reading 1 · L10

1. What is here too often?
2. What is missing?
3. Where does it break?
4. **What did the artist fight?**

REVISIT

Click to open



L10 · 17

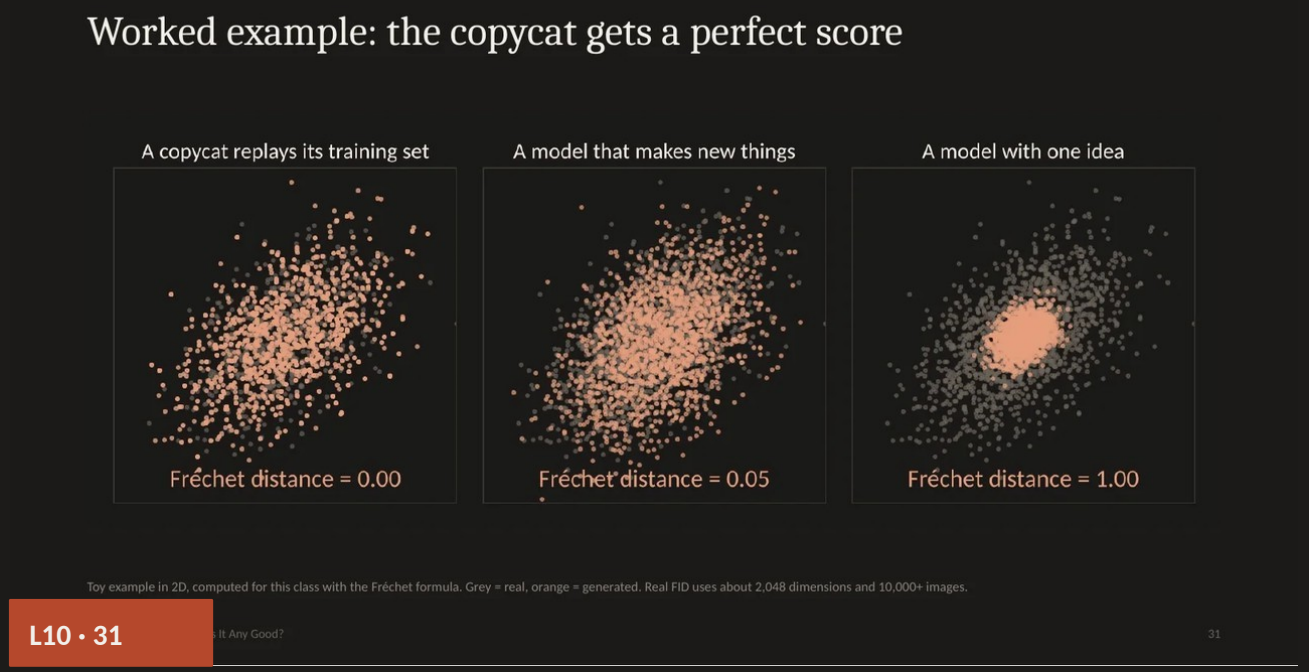


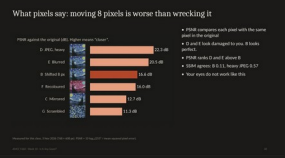
L10 · 18



L10 · 20

Measuring: what a number can see

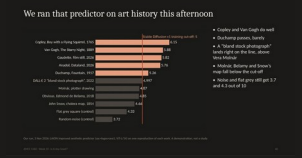




L10 · 28



L10 · 34



L10 · 40

Pixels and crowds

PSNR: an 8-pixel shift loses to a wrecked JPEG. FID: a copycat scores 0.00.

Sixteen numbers, or votes

VBench gives a profile; Elo's 43-point gap is nearly a coin flip.

Taste got into the model

LAION aesthetics chose SD's training data. Molnár: 4.87, below the line.

ASK YOURSELF

You generated 100 takes. Would you let a score pick your top ten?

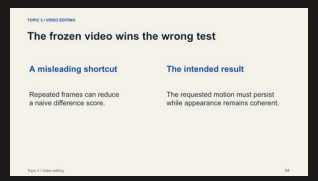
Every score we met, and its trap

Score	What it sees	Its trap
Frame difference · L02	Change between frames	The frozen video wins
PSNR · SSIM · L10	Pixel-by-pixel match	An 8-pixel shift scores worse than a wrecked JPEG
FID · L10	Distance between two crowds	A copycat scores a perfect 0.00
Sync-C / Sync-D · L08	Lip-audio offset	Wav2Lip “beats” real video: trained on the judge
Aesthetic score · L10	Predicted liking	Vera Molnár falls below the cut-off
Elo · L10	Blind A/B votes	A 43-point gap ≈ 56%: nearly a coin flip

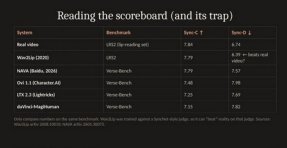
Goodhart's law: when a measure becomes a target, it stops being a good measure.

REVISIT

Click to open



L02 · 94

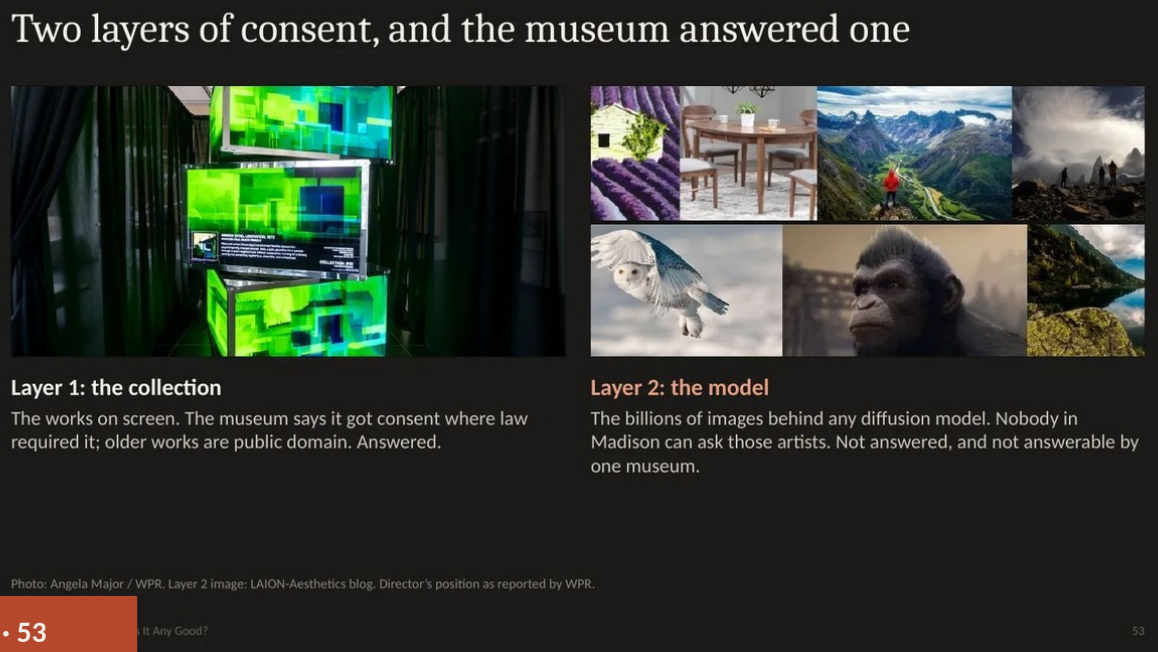


L08 · 44



L10 · 31

Showing: what happens when the work leaves your laptop





L10 · 48



L10 · 51



L10 · 55

Read the critics

Dataland: which of the four questions is each review answering?

Two layers of consent

Madison could answer for its collection, not for the model's training data.

Same word, different decisions

Molnár wrote the rule and could explain every line.

ASK YOURSELF

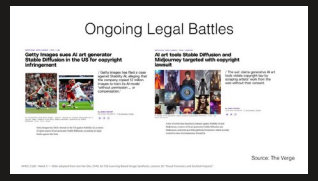
You run the Madison museum. Do you show Diffuse Control?

Consent has layers

Training data	L05 · L10	Billions of images behind the model. Nobody in one museum can ask those artists.
Likeness	L08	Voice and face. SAG-AFTRA 2026: informed consent and fair pay for replicas.
Places and bodies	L09	Who owns a scan of a place? Whose homes become robot training data?
The work itself	L10	Rights to the source material on screen: the layer Madison could answer.

REVISIT

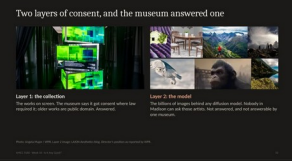
Click to open



L05 · 16



L08 · 88



L10 · 53

Sign your work, and don't apologise

1 **Visible label**

Promise: someone told you

2 **C2PA · Content Credentials**

Hash → claim → sign → embed
Tamper-evident, but stripped by a re-save. Absence proves nothing.

3 **Pixel watermark**

A hidden pattern, e.g. SynthID

CREDIT BLOCK · A2, A3, FINAL

- Title · length · medium
- Director: you
- Models + versions (pinned)
- Human work, counted: 41 attempts, 6 selected, edit, grade
- References: yours, or credited
- Consent: say it, even when “no real person”
- One plain sentence of disclosure

REVISIT

Click to open



L10 · 64

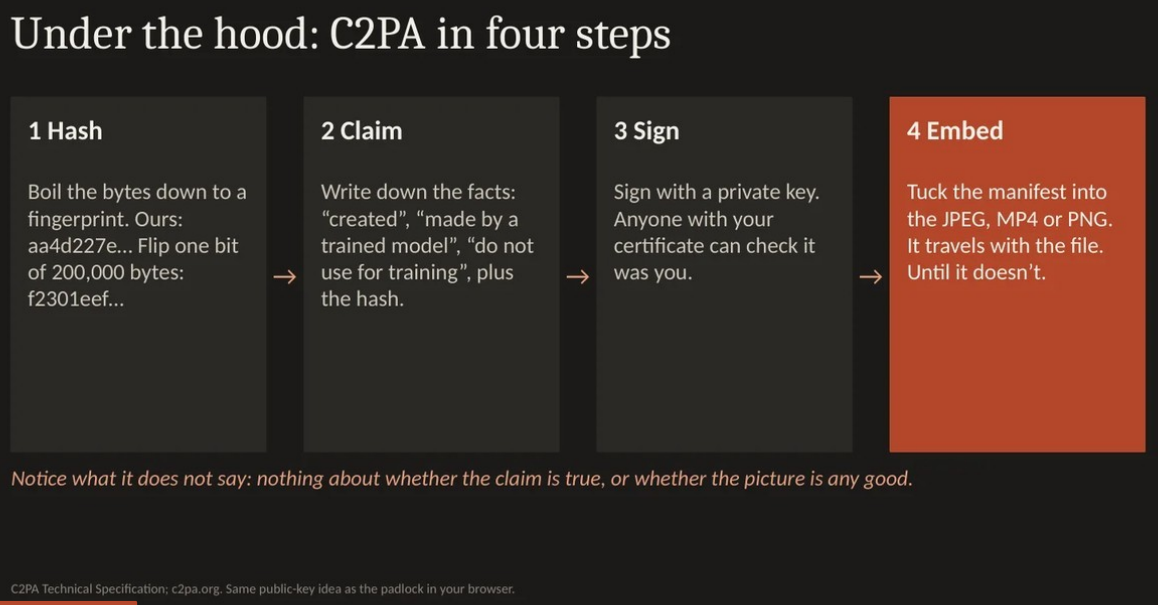


L10 · 67



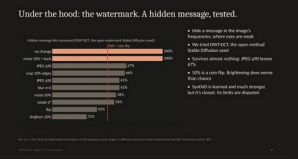
L10 · 73

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





L10 · 67



L10 · 68



L10 · 66

Hash → claim → sign → embed

It proves who said what about this exact file, not that it's true.

Tamper-evident, easily stripped

Re-save, screenshot, resize: gone. Edit with manifest: invalid.

Watermark, tested

DWT-DCT: 100% → 67% after JPEG (coin flip = 50%). SynthID is learned.

ASK YOURSELF

Absence of a credential proves nothing. Why?

Bring the semester into your final presentation

Explain a pipeline
L01 · L03

Your own four-stage diagram, with real model names and versions

Direct a coherent work
L01 · L02 · L04

Decisions, not outputs: your edit contract or beat sheet

Document an experiment
L01 · L03

One variable at a time; attempts counted; side by side

Critique with evidence
L02 · L10

Four steps on your own work; what did you fight?

Account for AI use
L07 · L08 · L10

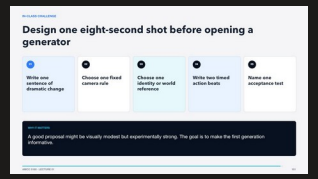
Credit block, pinned versions, consent line

Present a refined project
L04

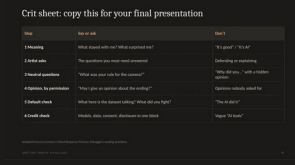
One clear intention, and what you would do next

REVISIT

Click to open



L01 · 101



L10 · 81



L10 · 73

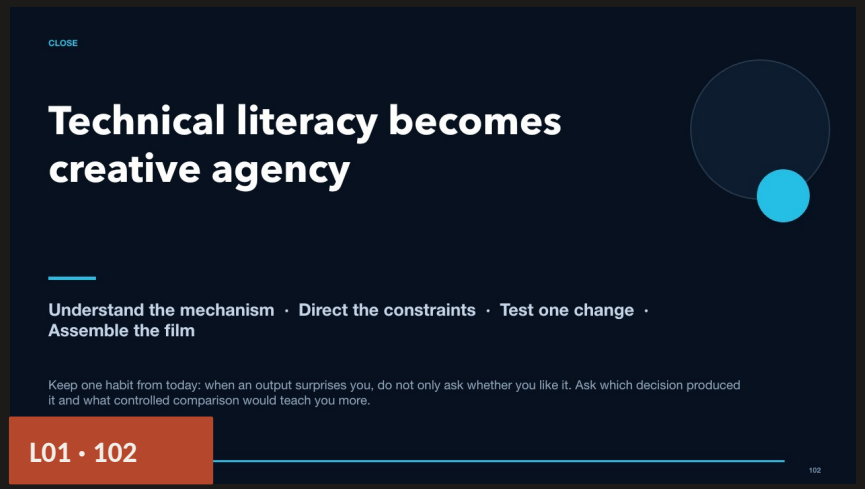
Three weeks left



Nov 10	Nov 17	Nov 24	Nov 30	Semester end
Tonight	**Finals I**	**Finals II**	**A2 + A3**	**Quizzes**
Last weekly presentations, after the break	Final presentations, sign-up page	Final presentations, sign-up page	23:59 HKT: both assignments due	Quizzes 01–11, editable until then

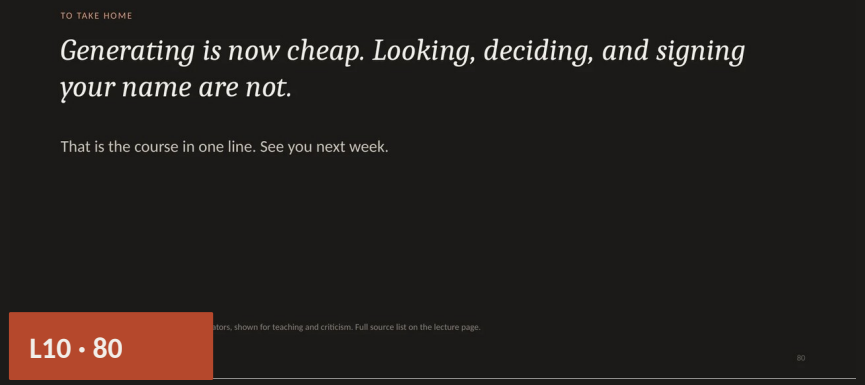
This week

Check your final slot and add your slides link · Revisit any slide from tonight you could not explain · Start Quiz 11 (review: each question connects two lectures)



LECTURE 01

When an output surprises you, ask which decision produced it, and what controlled comparison would teach you more.



LECTURE 10

Generating is now cheap. Looking, deciding, and signing your name are not.

Thank you. Ten-minute break, then the last weekly presentations.