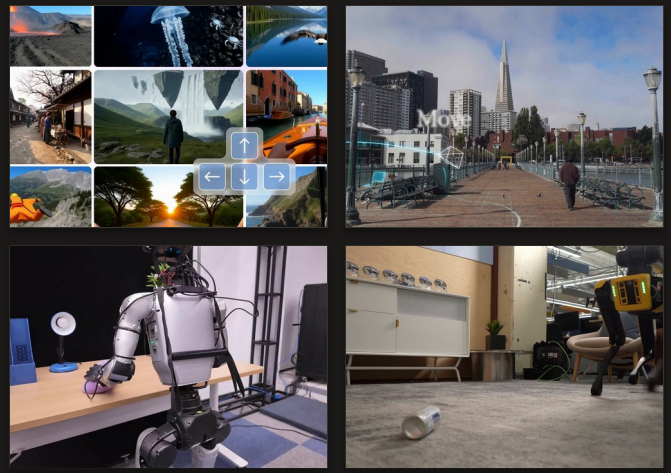
AMCC 5160 · LECTURE 09 · 27 OCTOBER 2026

Worlds You Can Walk Into,
Robots That Learn by Watching



World models, 3D, and robotics · Fall 2026

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

WARM-UP · 1 MINUTE

“I push this cup off the edge of the table. What happens next?”

Say it out loud: where does it land, what sound does it make?

Now: what if the floor is carpet? What if I push gently?

You just ran a world model in your head, conditioned on an action

Four stories from this year

01

From video to playable worlds

Genie 3, Project Genie and the Waymo World Model: video you can steer, one frame at a time

Genie 3 (2025) → Gemini Omni (May 2026)

02

3D is back: worlds as splats

Marble, Atlas and the open HY-World 2.0: generated worlds you can export, and an $8.2B bet on them

Atlas, 1 Sept 2026 · AMD × World Labs, 28 Sept

03

Robots that learn by watching

λ0 learns from 500 hours of first-person human video; Gemini 4 Argon turns a video into a simulator

Gemini 4 Argon, 13 Oct · λ0, 26 Oct 2026

04

Does a world model need pixels?

NVIDIA Cosmos 3 generates the world; LeCun's AMI Labs predicts it in latent space. Debate

Cosmos 3 (June) vs. JEPA

Where we've been: the pieces you already have

03

Lecture 03

World models: "what is being predicted?" The Astra round trip. VLA: see, read, act.

07

Lecture 07

Helios: a minute of video in real time. KV cache and distillation. A humanoid learns K-pop from phone video.

08

Lecture 08

Real-time characters: the latency budget, streaming distillation, avatars that listen.

→

Tonight: worlds

Put an action into the loop: real-time video becomes a world

→

Tonight: 3D

Keep the geometry: the world becomes a file you can open

→

Tonight: bodies

Close the loop with a real robot: the world becomes training data

You have seen two slides of tonight already

World models: what is being predicted?

Representation	Example output
State	Next position or a learned numerical state
Pixels	Next video frames conditioned on actions
3D scene	A scene representation that can be rendered
Action condition	The proposed movement affects the prediction

The representation determines which questions the model can answer.

83 AMCC5160 / Lecture 03

Lecture 03, slide 83: World models, what is being predicted?

VLA: See the scene, read the task, act

Vision + language → robot actions.

Gemini Robotics 2 adds whole-body control.

Google DeepMind examples • Gemini Robotics 2 • 30 July 2026

86 AMCC5160 / Lecture 03

Lecture 03, slide 86: VLA, see the scene, read the task, act

Tonight opens both boxes: how the prediction is made, and how a robot learns the action.

How a top course maps the field: three paradigms

Three Paradigms for World Models

1. Generate a 3D Scene, then Render/Simulate

Examples:

- Marble (World Labs)
- GSGen (BNRist)
- DreamFusion (Google Research)
- NeRF-VAE (DeepMind)

Creates a representation that can be rendered using an existing renderer

2. Interactive Videos

Examples:

- Genie 1-3 (DeepMind)
- GameNGen
- Muse (Microsoft)
- Oasis (Decart AI)
- GAIA 1-2 (Wayve)

The feel of interacting with a game engine

3. Latent World Representations

Examples:

- PAN (MBZUAI)
- V-JEPA 1-2 (Meta)

Represent the world state in a low dimensional latent space (continuous, discrete, or both)

13

THE TAKEAWAY

1 · Generate a 3D scene, then render it (Marble, GSGen) → our Story 2

2 · Interactive video (Genie, GameNGen) → Story 1

3 · Latent world representations (V-JEPA) → Story 4

Robots (Story 3) borrow from all three

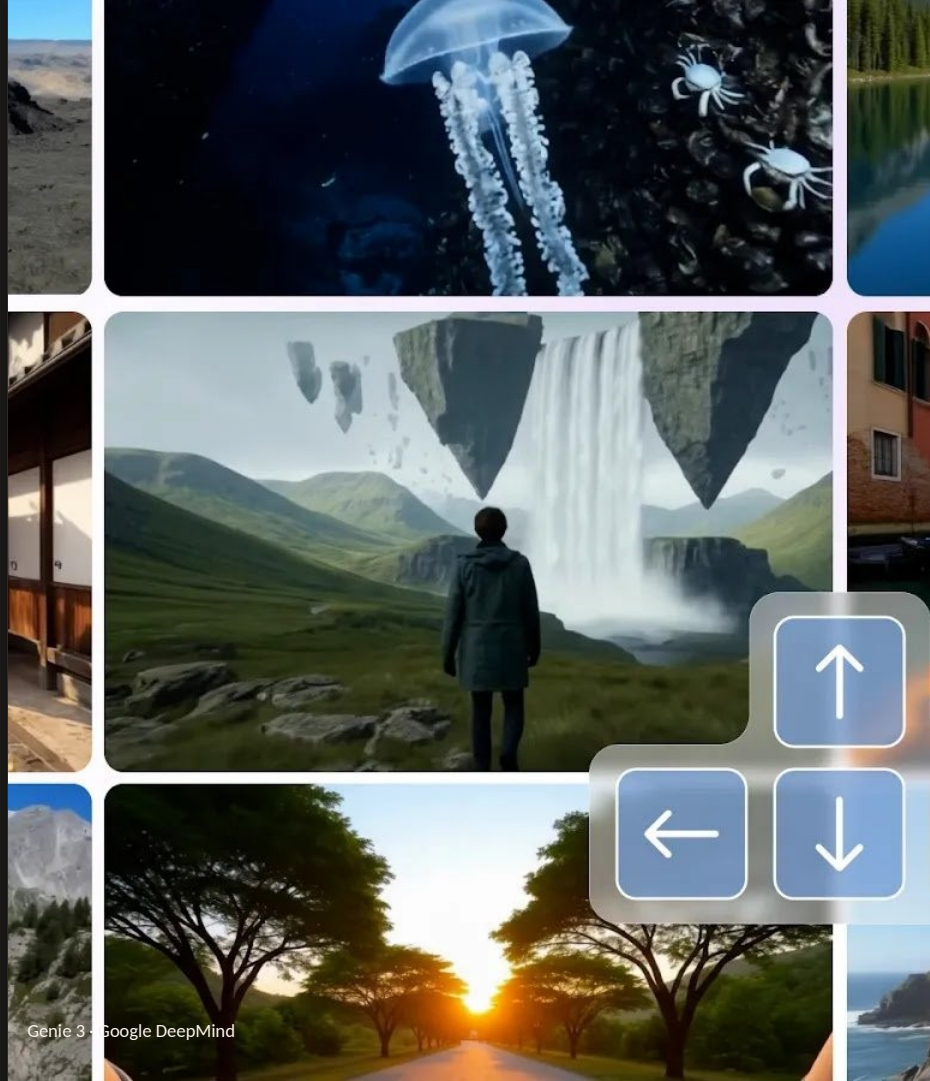
Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

STORY 1 OF 4

01

From video to playable worlds

Press a key, and the model draws the next moment



What is a world model?

What is a World Model (WM)?

the psychology literature has long defined our notion of a world model as “hypothetical thinking”

informally, we speak about this as a “thought experiment”

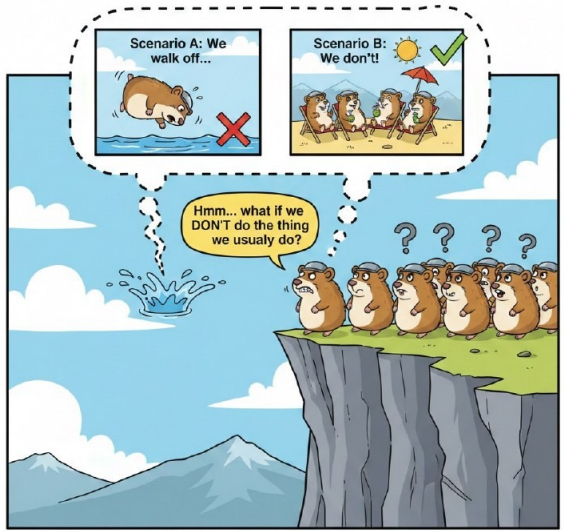


Figure generated by Nano Banana

9

DEFINITION

A model that takes the previous world state and an action, and predicts the next state

$p(s' \mid s, a)$

The “state” can be pixels, 3D, or an abstract vector

Video model: what happens next. World model: what happens next if I do this

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

Eight years, from a dream of a car race to a playable world

① **2018 · Ha & Schmidhuber**

"World Models": an agent learns to drive inside its own dream of a racing game

② **2024 · GameNGen**

A diffusion model runs DOOM at about 20 fps, with no game engine

③ **Feb 2024 · Genie**

Learns playable 2D platformers from unlabelled internet video, ~1 fps

④ **Dec 2024 · Genie 2**

3D worlds from one image, 10–20 seconds, 360p

⑤ **Aug 2025 · Genie 3**

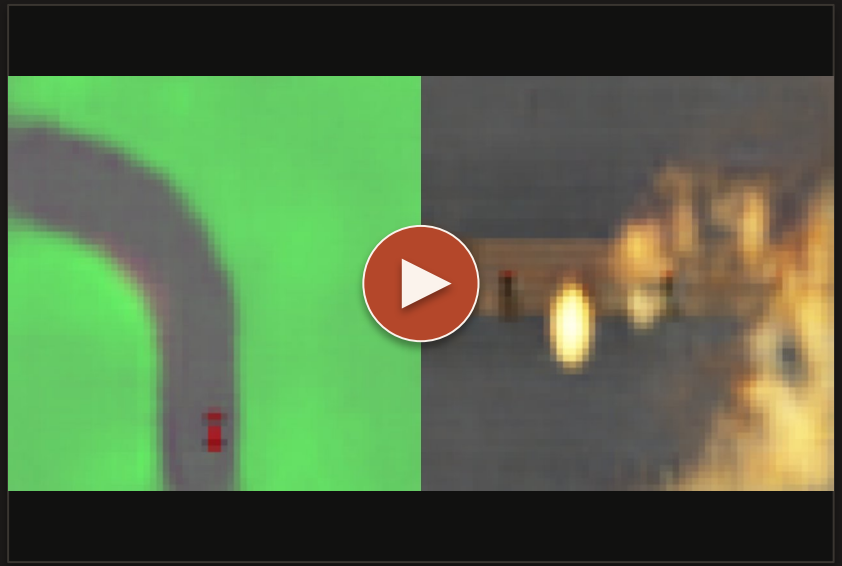
Real time: 720p, 24 fps, a few minutes of consistency

⑥ **2026 · Project Genie, Gemini Omni**

A public prototype (60-second sessions) and a "world model" as a consumer product

Ha & Schmidhuber 2018 (worldmodels.github.io); Valevski et al. 2024 (GameNGen); Bruce et al. 2024 (Genie); Google DeepMind Genie 2 and Genie 3 posts; Google I/O, 19 May 2026 (Gemini Omni).

Two ancestors: a dream of a race, and DOOM without an engine



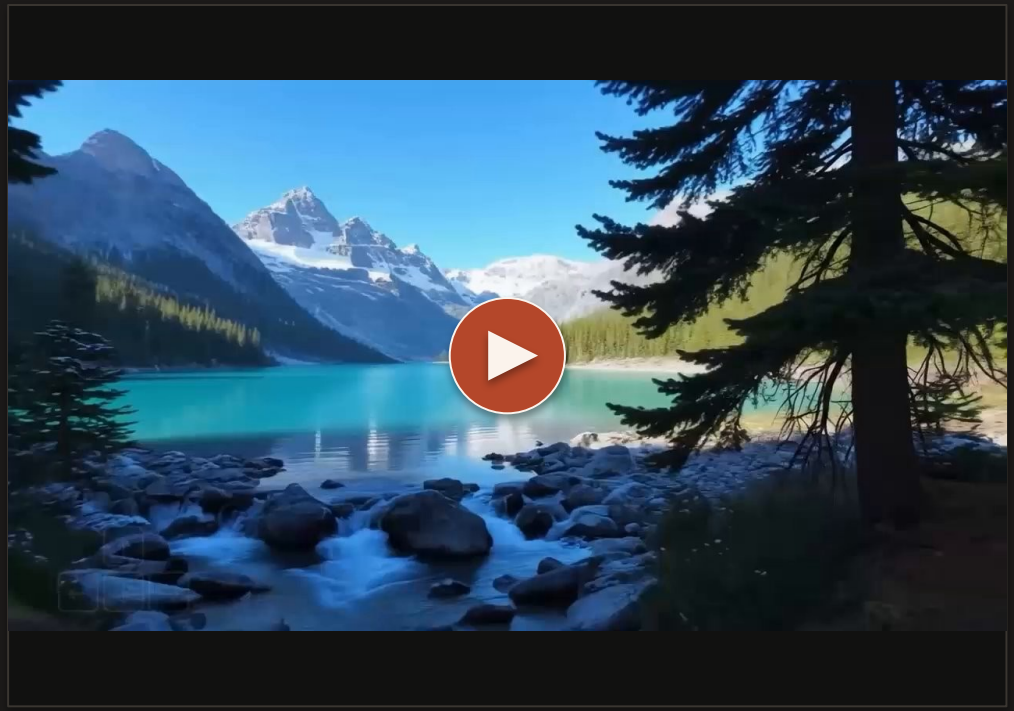
Ha & Schmidhuber (2018): the real frame vs. the agent's compressed "dream"

GameNGen (2024): every DOOM frame is generated, live, by a diffusion model

Left: learn a small world and practise inside it. Right: make the world itself out of a generator. Genie joins the two.

Ha & Schmidhuber, "World Models", 2018 (worldmodels.github.io); Valevski et al., "Diffusion Models Are Real-Time Game Engines", 2024 (gamengen.github.io).

Genie 3: a world that waits for you

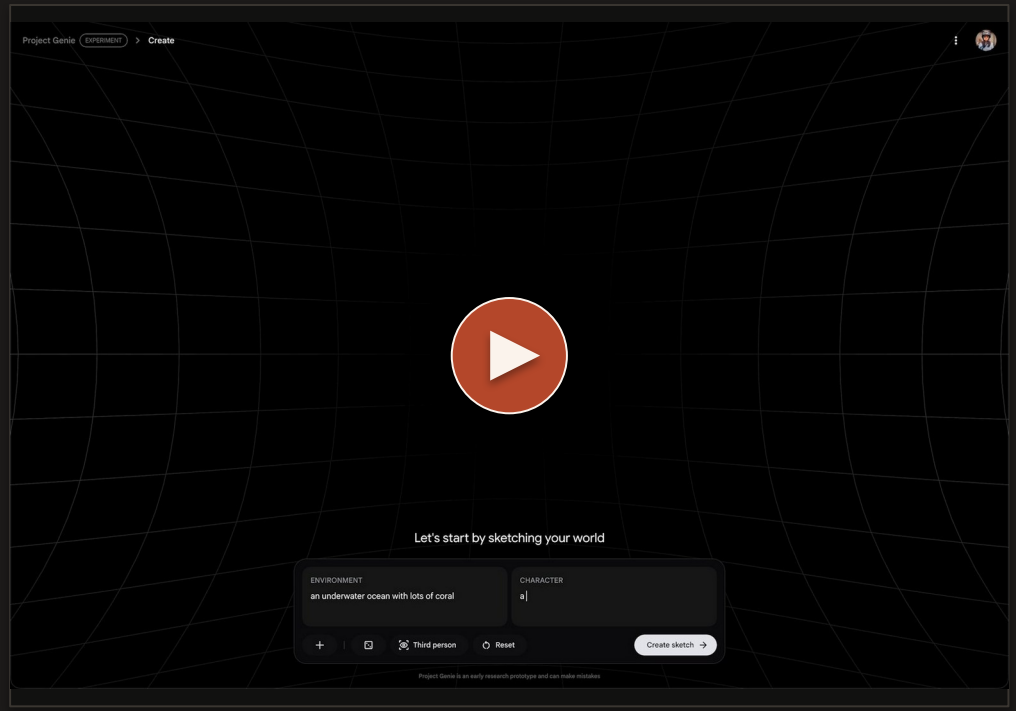


- Text prompt → explorable world
- 720p, 24 fps, real time
- Remembers what it drew for about a minute
- “Promptable world events”: type “a storm starts” mid-walk

▶ Genie 3 demos

Google DeepMind, “Genie 3: A new frontier for world models”, August 2025.

Project Genie: anyone (with a subscription) can try it

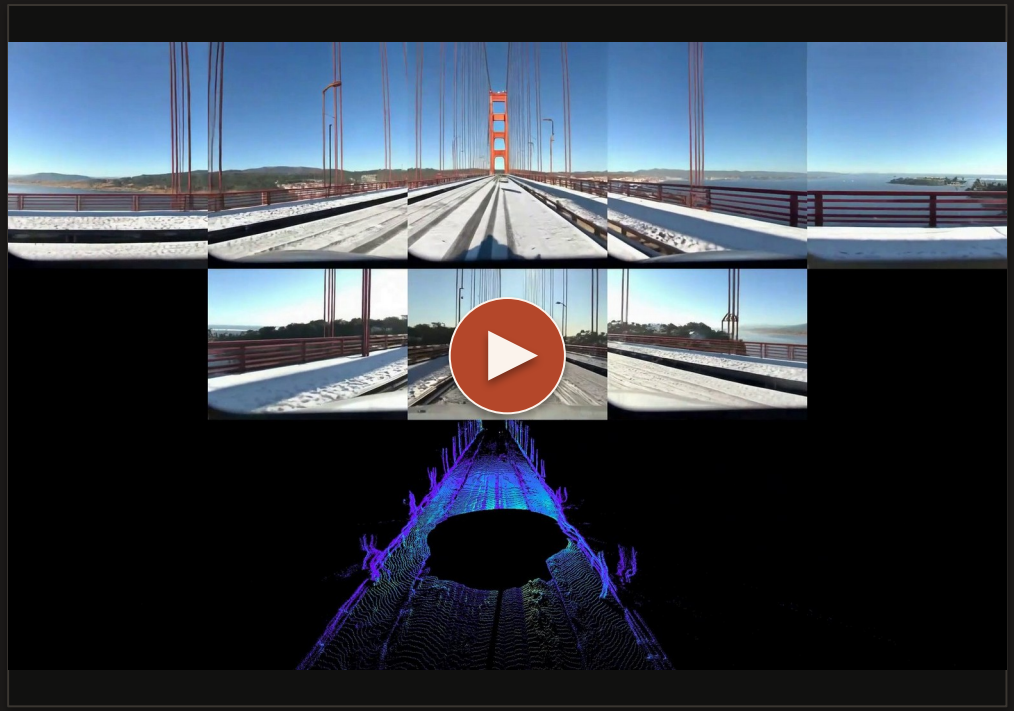


- A prototype combining Genie 3, Nano Banana Pro and Gemini
- For Google AI Ultra subscribers in the US
- Sessions capped at 60 seconds: compute is the limit
- Google says it is also collecting feedback and data

▶ **Project Genie**

Google, “Introducing Project Genie”, 29 January 2026.

Waymo World Model: Genie 3 learns to drive



- A Genie 3 variant built for self-driving
- Generates camera views and lidar
- Controls: the car's actions, scene layout, other drivers, weather
- Used to rehearse rare, dangerous edge cases

▶ Waymo World Model

Waymo, "The Waymo World Model: a new frontier for autonomous driving simulation", February 2026.

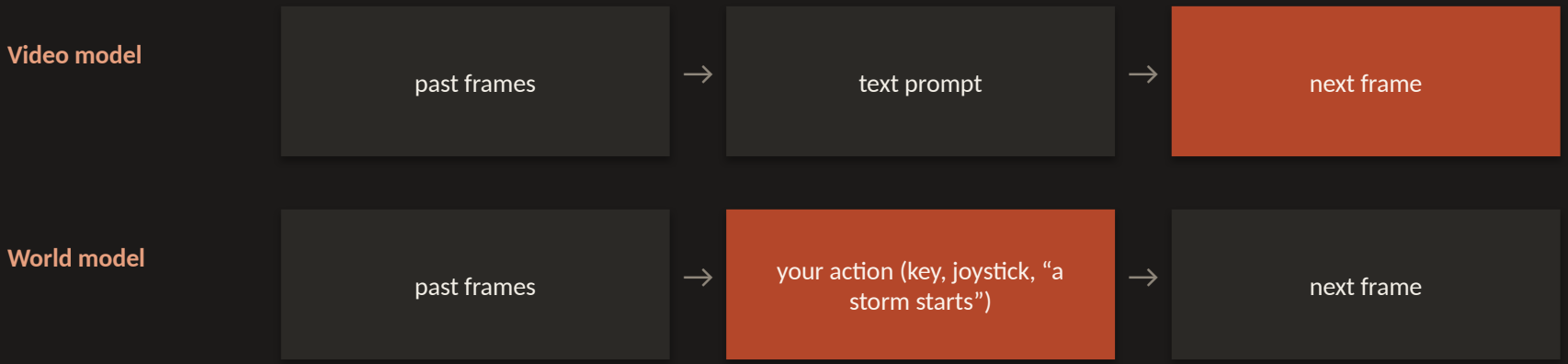
Gemini Omni: “world model” becomes a product name

- Announced at Google I/O, 19 May 2026
- Any input (text, audio, image, video) → video
- Edit by conversation: camera, background, style
- Every output carries a SynthID watermark

▶ Gemini Omni

Google I/O 2026; Yahoo Tech, “Gemini Omni is Google’s new world model”, May 2026.

Under the hood: one extra input



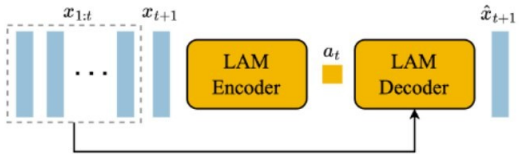
video: $p(x_{t+1} \mid x_{\le t}, \text{text})$ world model: $p(x_{t+1} \mid x_{\le t}, a_t)$

The architecture can be almost the same autoregressive video model from Lecture 07. The difference: an action enters at every step, so the model must generate fast enough to wait for you.

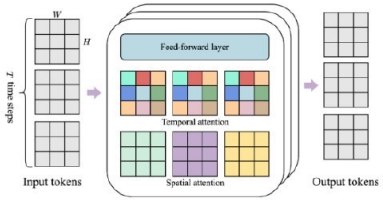
Under the hood: where do the actions come from?

Genie-1 Model

Latent Action Model:



ST-Transformer:



Latent Action Model (LAM):

- An unsupervised model for learning *latent* actions from unlabled videos
- Another VQ-VAE model
  - Encoder: converts t frames of video plus the t+1 frame ($\mathbf{x}_{1:t+1}$) into sequence of t+1 actions ($\mathbf{a}_{1:t+1}$)
  - Decoder: converts t frames ($\mathbf{x}_{1:t}$) and t actions ($\mathbf{a}_{1:t}$) into the t+1 frame ($\mathbf{x}_{t+1}$)
- At test time, this model is discarded since actions come from a human user

- *Backbone model:* ST-Transformer

Figure from http://arxiv.org/abs/2402.15391

28

GENIE'S TRICK

Internet video has no key presses

Genie learns a "latent action" between two frames: one of 8 discrete codes

Trained without any labels

At play time, your key press picks a code

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

Genie at play time: three parts in a loop

Genie-1 at Test Time

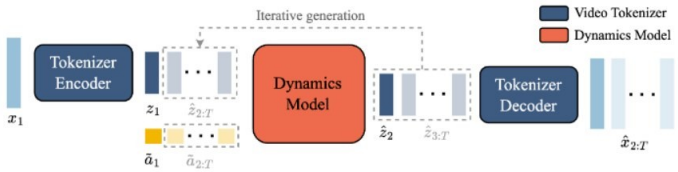


Figure 8 | **Genie Inference**: the prompt frame is tokenized, combined with the latent action taken by the user, and passed to the dynamics model for iterative generation. The predicted frame tokens are then decoded back to image space via the tokenizer's decoder.

Figure from http://arxiv.org/abs/2402.15391

30

THE LOOP

1 · Tokenizer: frame → tokens

2 · Your action → a latent action code

3 · Dynamics model predicts the next frame's tokens

Decode → show → repeat, every frame

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

GAME · 3 MINUTES

Be the latent action model

I show two frames from a game you have never played

You have only 8 hand signals. Agree as a row: which one explains the change?

Then: did every row invent the same 8 “actions”?

Under the hood: why the world must come back the same

Walk away	see a red door	→	turn left 10 s	→	turn back
Pass	the red door is still there	→	same place, same paint	→	“object permanence”
Fail	the door is blue, or gone	→	the room has re-arranged	→	drift (Lecture 07)

Lecture 03’s round-trip test, now in real time. Genie 3 keeps visual memory for about a minute; long-take video in Lecture 07 drifted for the same reason.

Under the hood: the real-time budget

24 fps

the target: one new frame every 41.7 ms

→

~40 ms

to read your key, run the model, decode and display

→

KV cache

reuse the past instead of recomputing it (Lecture 07)

→

Few steps

distil 50 denoising steps into a handful (Lecture 07, 08)

The same three tricks made Helios real-time in Lecture 07 and talking avatars real-time in Lecture 08. A world model is the same budget, with your hand on the controls.

Genie 3 vs. Genie 2 vs. GameNGen

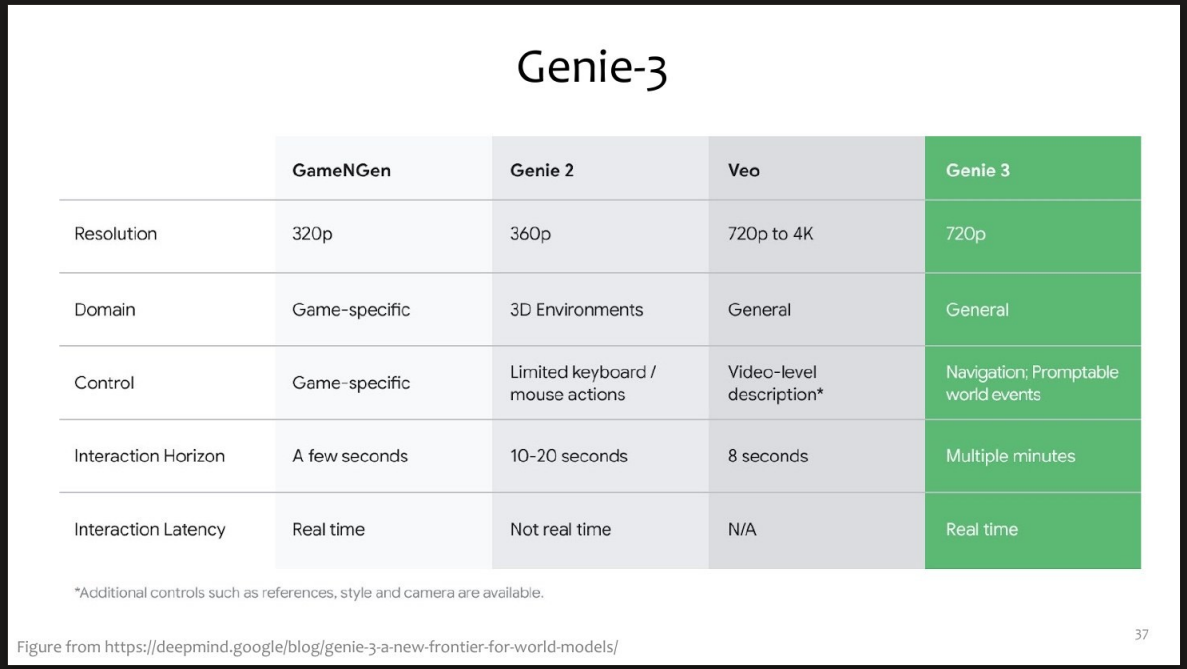


Figure from https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

READ THE TABLE

Resolution: 360p → 720p

Domain: one game → any 3D world

Interaction horizon: seconds → minutes

Latency: real time, at 24 fps

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

What world models still get wrong

Genie 3 Limitations

- Interaction window is limited to a short duration (a few minutes)
- Actions (are likely) latent and come from a fixed set; defining new actions is not possible
- Prompting for changes in the world is possible and represents an open set; thus representing an asymmetry between the closed action set and the open world change set
- Limited to one character (at least that seems to be the case)
- Fundamentally just a video model: no game-engine-interpretable representation of the world (e.g. point cloud, mesh, Gaussian splatting) is provided
- Super high computational demands

39

LIMITS (CMU)

Interaction lasts minutes, not hours

Action space is limited and fixed

Several characters interacting: hard

Text in the world is unreliable

It is video: no mesh, no 3D file to export

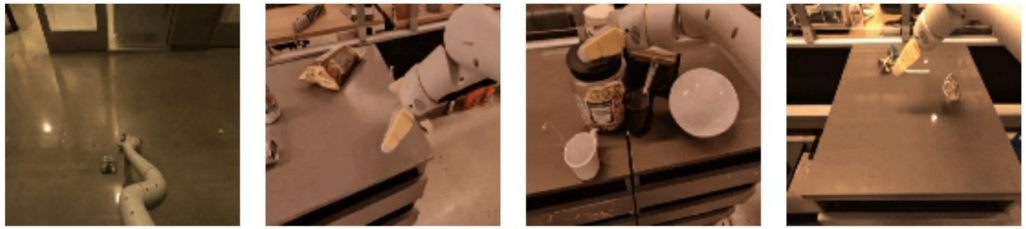
Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

A bridge to robots: Genie dreams of a robot arm

Genie-1 Application to Robotics

- Robotics dataset:
  - 130k robotics demonstrations
  - a simulation dataset (size unspecified)
  - 209k episodes of real robot data
- No actions included, just videos

- Training Genie on the robotics datasets learns a new model that can simulate "game" style play with a robotic arm as shown below



latent actions: 0, 0, 1, 2, 6, 1, 1, 4, 0, 0

Figure from https://sites.google.com/view/genie-2024/home

32

ALREADY IN 2024

Same recipe on ~130k robot videos, no action labels

The latent actions became robot-arm moves

A world model is a place a robot can practise

→ Story 3

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

Useful: directing a world, not a clip

1

World

Place, era, weather, lighting. “A flooded night market in Mong Kok, neon reflections”

2

Body

Who are you? Height, speed, camera: first person, drone, a cat

3

Events

What can happen while you walk: “a tram passes”, “the rain stops”

4

Test

Walk away and come back; try to break it; note where it drifts

Your Assignment 3 game: use a world model as a level sketch, then rebuild what works.

DISCUSSION · 3 MINUTES

If a world only exists while someone is playing it, what is the artwork: the world, the play, or the recording?

STORY 2 OF 4

02

3D is back: worlds as Gaussian splats

Generated worlds you can keep, export and walk a robot through

Pixels or geometry? Two ways to make a world

A

Draw the next frame (Story 1)

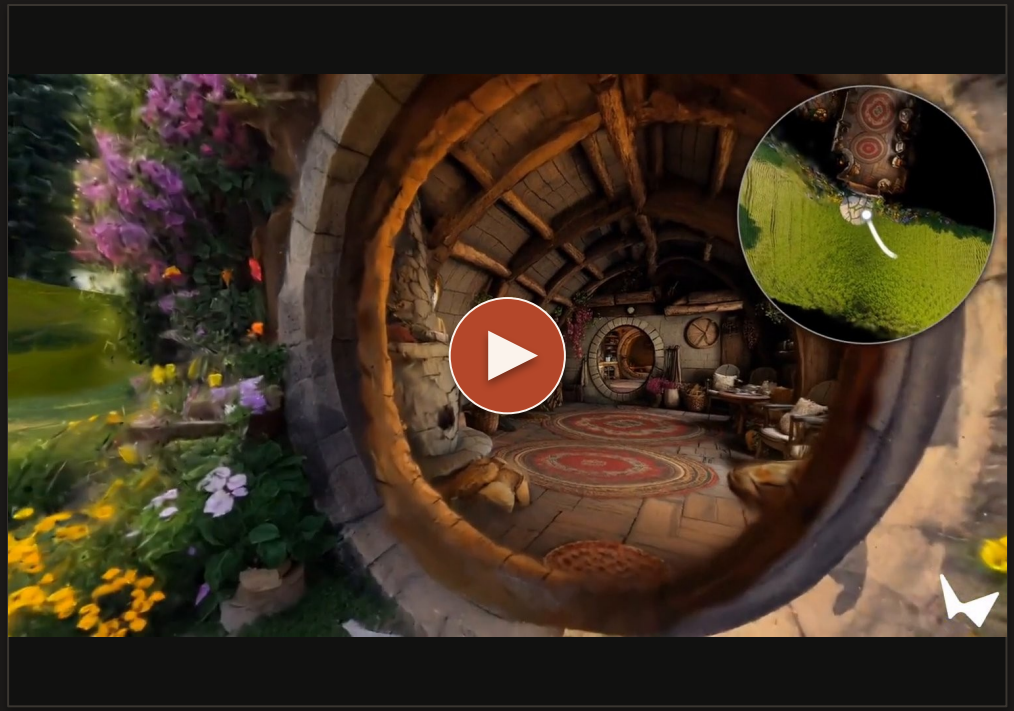
Genie 3, Gemini Omni
+ Anything can happen: water, crowds, storms
+ Feels alive
– Forgets after about a minute
– Nothing to export; every frame costs a GPU

B

Build the scene, then render (Story 2)

Marble, Atlas, HY-World 2.0
+ Consistent forever: it is a file
+ Open it in Blender, Unity, a headset, a robot simulator
– Mostly static: things don’t move by themselves
– Rendering is cheap, generation is not

Marble: one picture in, a world you can walk around



- Input: text, one image, several images, a 360° panorama, or video
- Output: a Gaussian-splat world (.spz / .ply)
- Marble 1.1 (April 2026): better light, auto-expanding worlds
- View on desktop, phone or VR with the open Spark renderer

▶ Marble 1.1

World Labs, Marble 1.1 release notes and “Generating bigger and better worlds”, April 2026.

Atlas (1 Sept 2026): one model for pictures, video and 3D

- Everything goes into one “spatial context”: each image placed at a camera pose in 3D
- Outputs: new views, minute-long 1440p video, depth, point clouds, splats
- Two or three phone photos → a faithful 3D scene
- Early access only; will power future Marble versions

▶️ **Atlas demos**

World Labs, “Atlas”, 1 September 2026. worldlabs.ai/blog/atlas

Atlas, by World Labs' own numbers

75–94%

of raters preferred Atlas for camera-controlled video, vs. five video models incl. Seedance 2.5 and Gemini Omni Flash

2–3

photos are usually enough for a faithful reconstruction; it can hold 100+

24

phone-video frames per room to simulate a robot driving through it

25.3

average 3D error ($\times 10^{-3}$), vs. 28.7–47.7 for the best open reconstruction models

Company numbers, no paper yet. Treat them as claims, and test them: walk away and come back.

The money: a chip company buys a world-model company

$1B

World Labs' funding round, February 2026, with AMD, NVIDIA and Autodesk

→

~$8.2B

AMD's all-stock deal to acquire World Labs, announced late September 2026

→

Fei-Fei Li

to become AMD executive VP and chief scientist

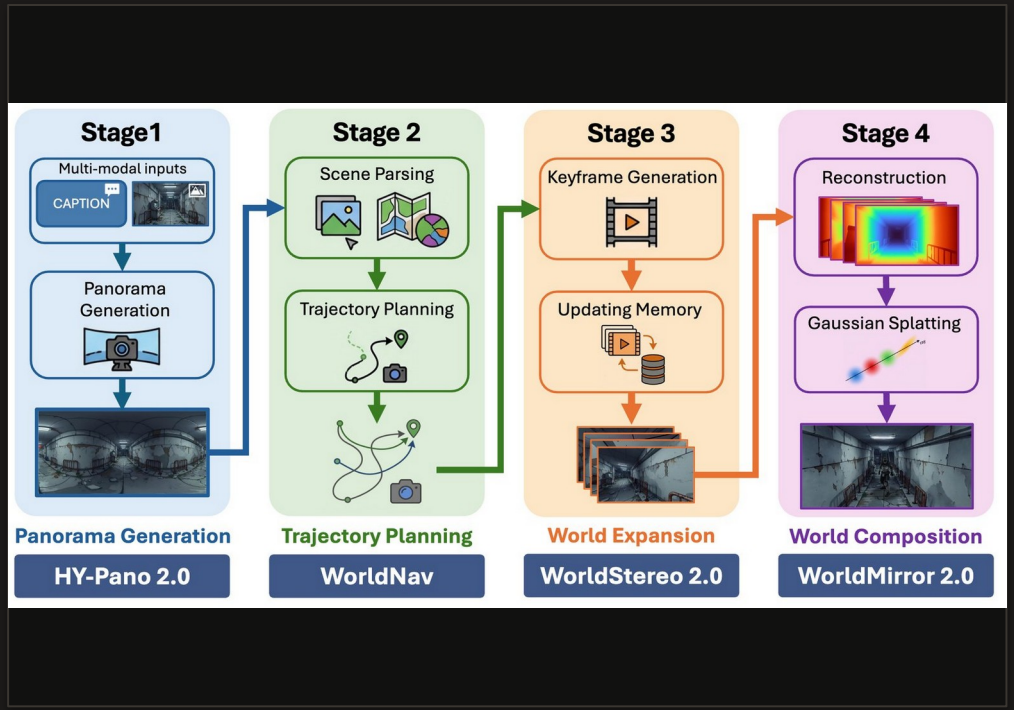
→

End 2026

expected closing, subject to regulators

Why would a chip maker buy a world-model lab? Every frame of a world, and every robot practising in it, is compute. Not yet closed: what happens to Marble and Spark is not announced.

The open alternative: Tencent HY-World 2.0

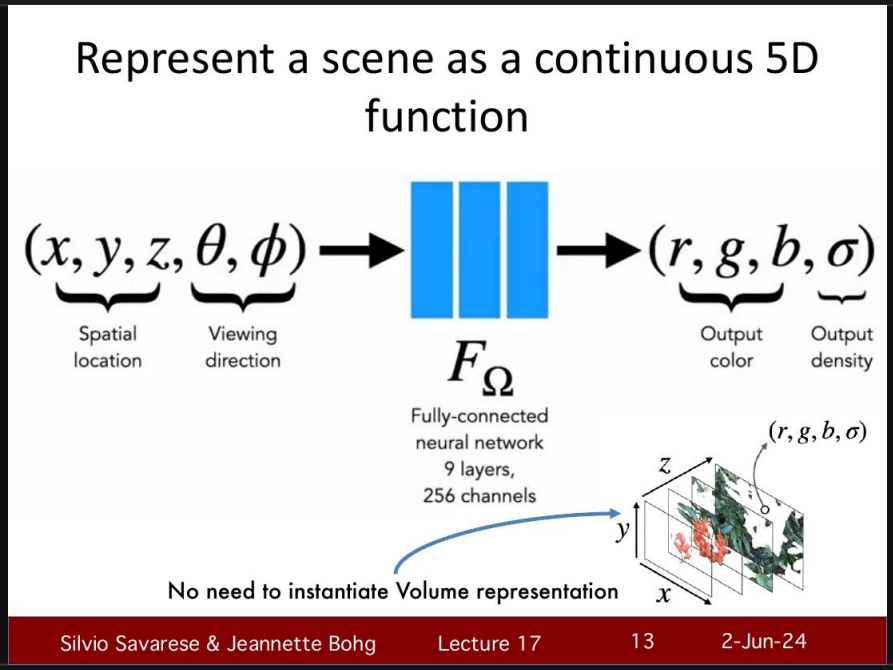


- Text or one image → a navigable 3D world; photos or video → a reconstruction
- Pipeline: panorama → camera path → expand views → 3D in one pass (WorldMirror 2.0)
- Exports mesh, 3D Gaussian splats, point clouds → Blender, Unity, Unreal, Isaac Sim
- Open code and weights released April–May 2026

▶ HY-World 2.0 project page

Tencent Hunyuan, HY-World 2.0 technical report (arXiv 2604.14268) and GitHub, April 2026.

Under the hood: from a scene to a function (NeRF)



NERF, 2020

A network answers: at point (x, y, z), seen from direction (θ, φ), what colour and how dense?

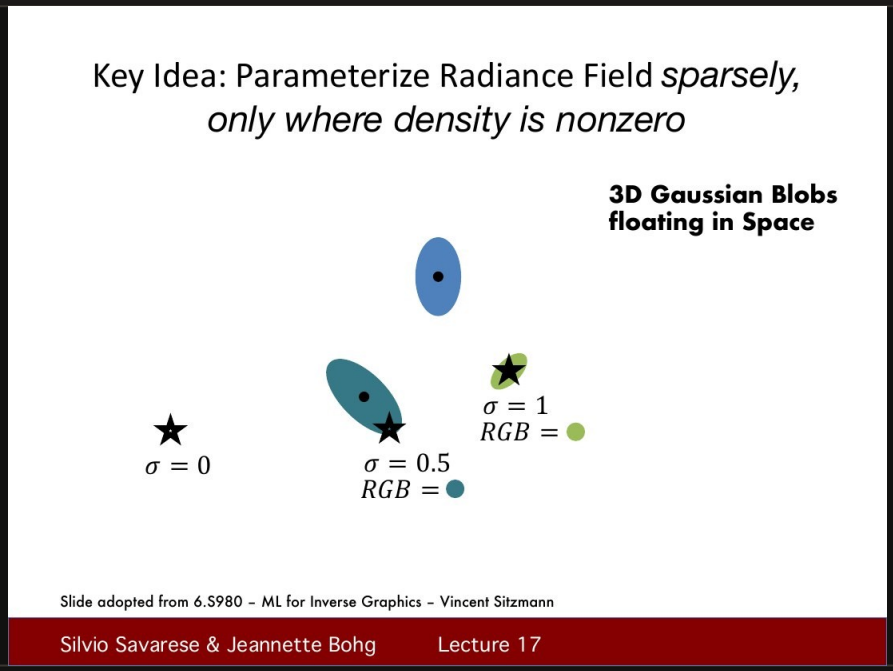
Render by marching along each camera ray

Train by matching your photos

Beautiful, but slow: ask the network millions of times per image

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

Under the hood: only put stuff where stuff is



3D GAUSSIAN SPLATTING, 2023

Most of space is empty air

So: millions of small 3D blobs (Gaussians), only where there is something

Each blob: position, shape, opacity, colour

“Splatting”: throw each blob onto the screen like a paint splat

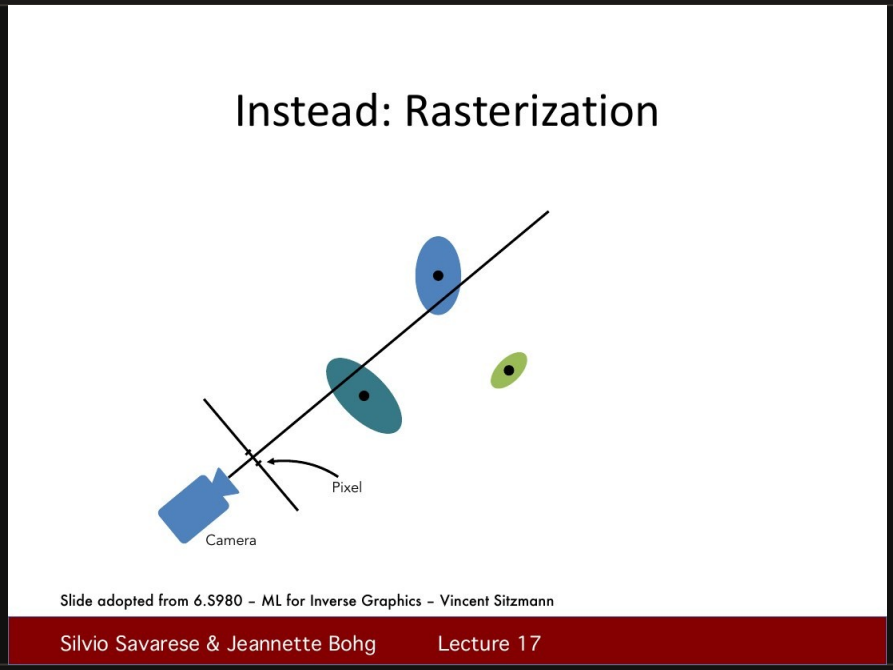
Adapted from Stanford CS231A Lecture 17 "Gaussian Splatting for Novel View Synthesis" (Silvio Savarese & Jeannette Bohg; slides adapted from MIT 6.S980, Vincent Sitzmann).

Worked example: what is inside one splat?

3	→	3 + 4	→	1	→	48	→	= 59
position: x, y, z		shape: scale on 3 axes + a rotation (quaternion)		opacity		colour that changes with viewing angle (spherical harmonics, 16 × RGB)		numbers per splat ≈ 236 bytes. 1 million splats ≈ 236 MB raw

So a room of 1–3 million splats is a few hundred MB as .ply. Compressed formats like .spz (used by Marble) are roughly 10× smaller, which is why splats now stream to phones.

Under the hood: rasterize, don’t ray-march



WHY IT IS FAST

Project every blob onto the screen as a 2D ellipse

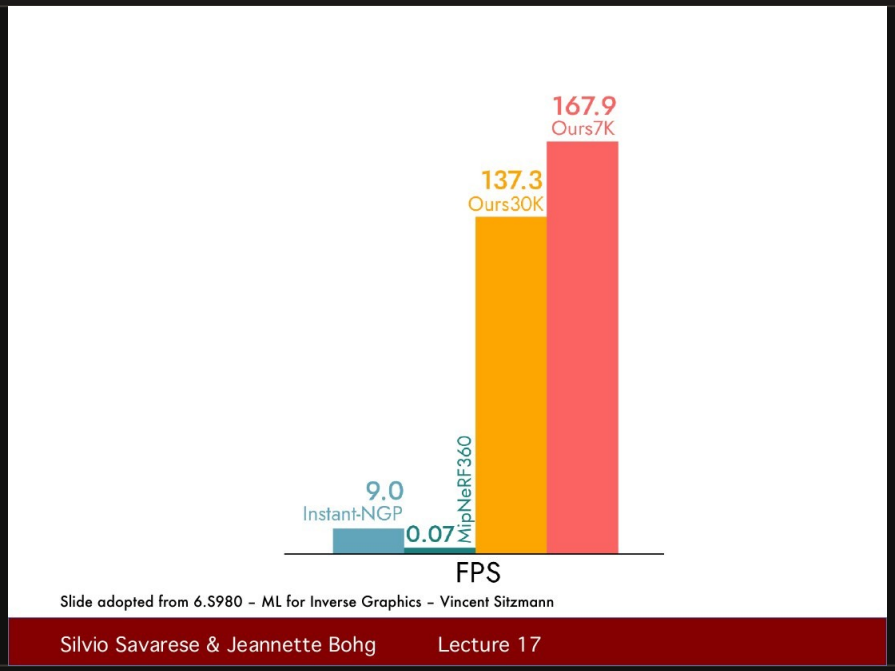
Sort by depth, blend front to back

This is what GPUs were built for (like games)

No neural network at render time

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

The result: from 0.07 to 100+ frames per second



SPEED

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

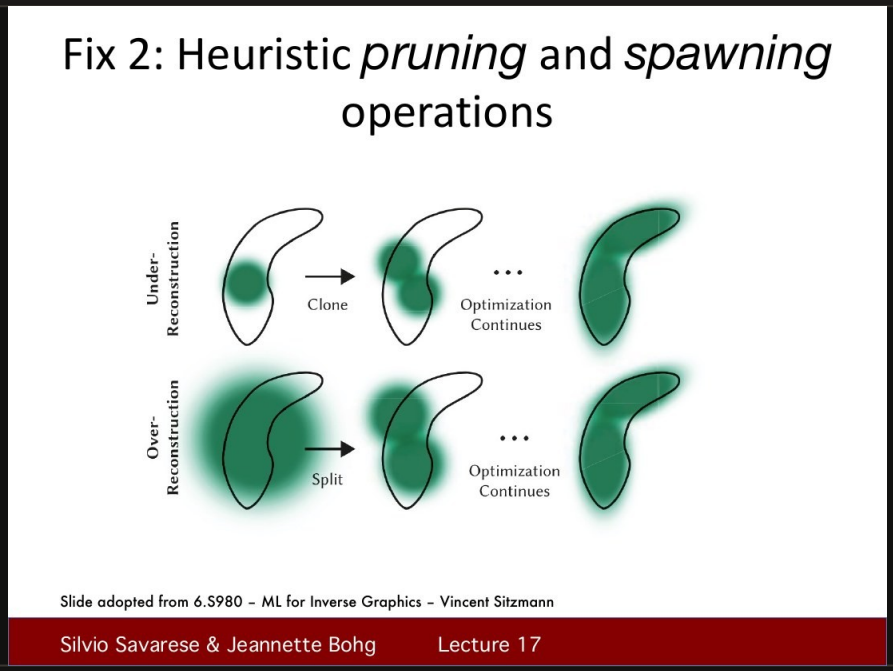
Instant-NGP: ~9 fps

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).

Under the hood: how splats are trained



TRAINING

Start from the 3D points that photo-matching (SfM) already found

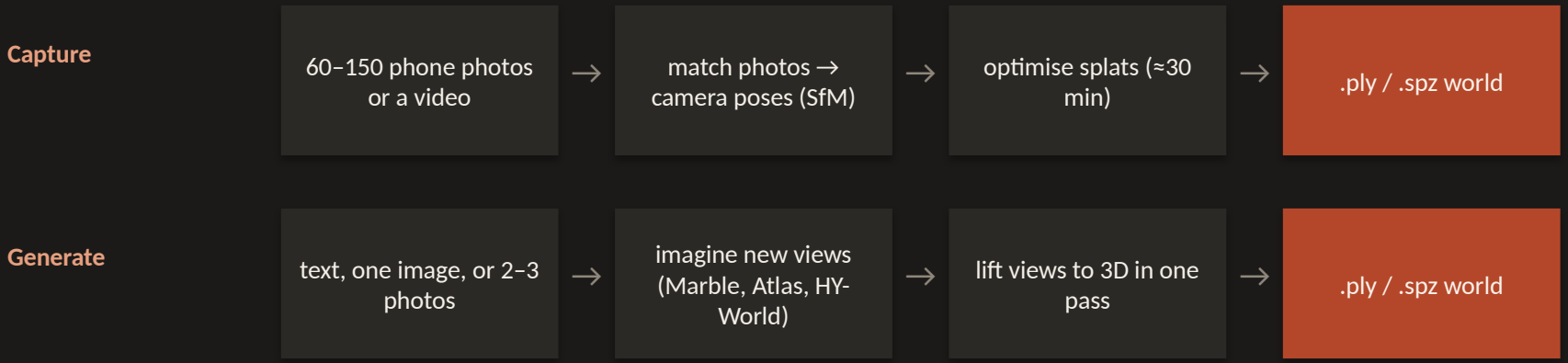
Render, compare to your photos, nudge every blob

Split blobs that are too big, clone where detail is missing, prune invisible ones

~20–40 minutes on one GPU for a room

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

Capture or generate: two roads to the same file



Capture is faithful but needs the place. Generation needs only a prompt but invents what it cannot see. Atlas blurs the line: more photos means less invention.

This autumn, splats became a normal file format

- Blender 5.3 (alpha): native PLY / SPZ splat import and render in EEVEE and Cycles
- three.js r186: a built-in splat renderer for the web
- Khronos: a glTF Gaussian-splatting extension
- SuperSplat 3.0 editor; 4D splats streamed for live sport

▶ **Radiance Fields: Sept 2026 round-up**

Radiance Fields, “Gaussian Splatting in September 2026”; Khronos Group glTF press release.

Useful: phone → splat → Blender, in one evening

1

Capture

Orbit slowly, 60–150 photos or a 1-minute video. Even light, no moving people, lots of overlap

2

Train

Free/desktop: Postshot, Brush, or nerfstudio's gsplat. Or an app (Polycam, Scaniverse)

3

Clean

SuperSplat in the browser: crop floaters and sky, set the floor level

4

Use

Blender 5.3 or three.js / Spark; or skip 1–2 and ask Marble or HY-World to generate

Tip: a splat is lit by the light you captured. Shoot at the hour you want the scene to feel like.

Three art prompts for a splat

I

A place that is about to disappear

Scan a shop, a room, a street before it changes. A splat freezes a place the way a photo freezes a moment

II

An impossible join

Give Atlas or Marble two unrelated photos and let it invent the corridor between them. What does the model think connects them?

III

A level for your game

Scan a real corner of campus, then break it: change the light, remove the floor, add one impossible door (Assignment 3)

DISCUSSION · 3 MINUTES

A photo freezes a moment. A splat freezes a place. What should you be allowed to scan, and who owns the copy?

Your own room?

A shop, a temple, a museum gallery?

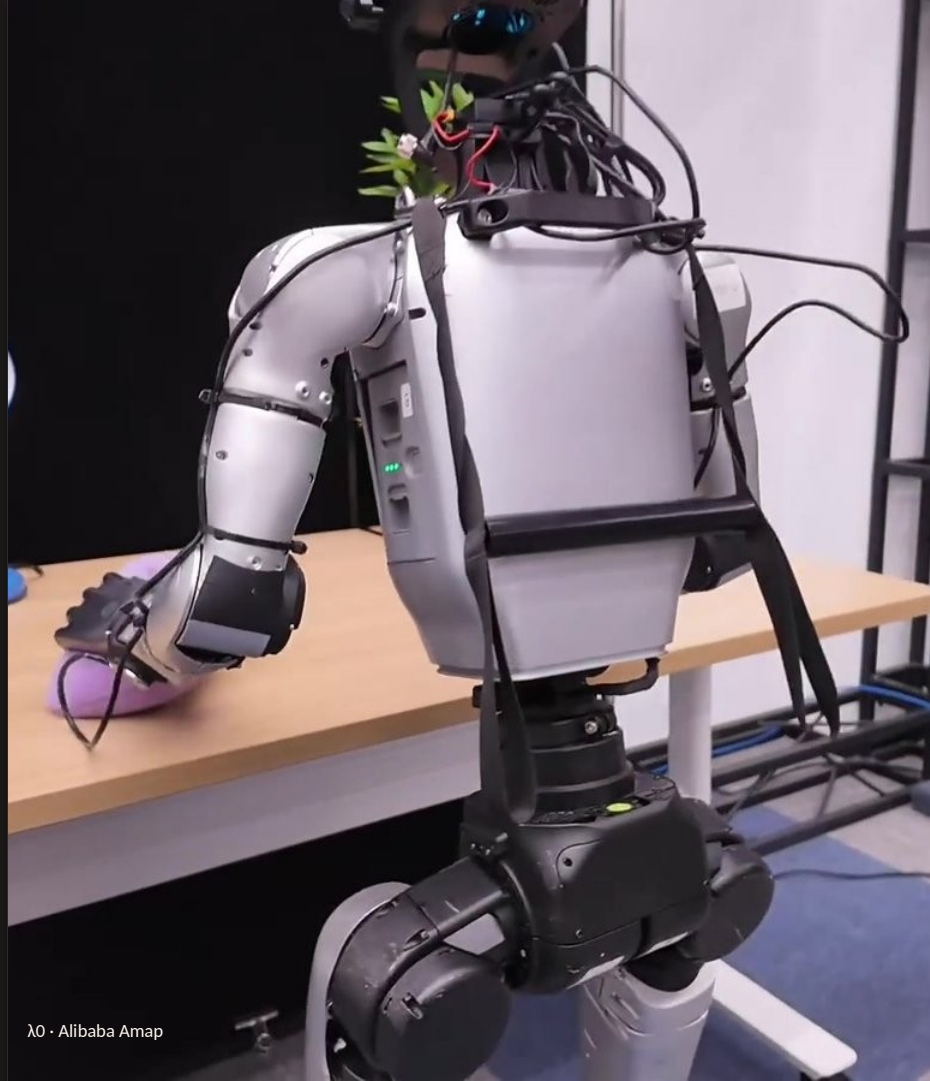
A stranger’s home, from photos they posted online?

STORY 3 OF 4

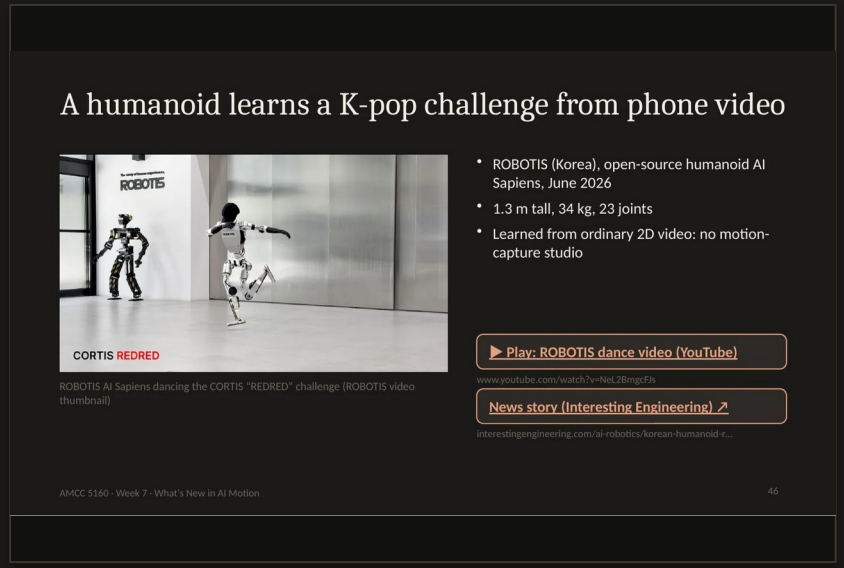
03

Robots that learn by watching

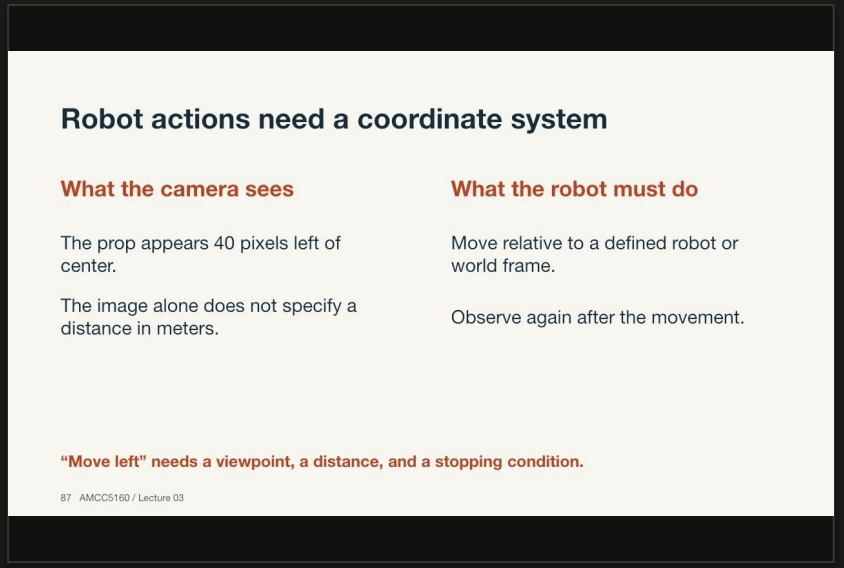
First-person video, a simulator built from a video, and a model released yesterday



Where we left robots: Lecture 07 and Lecture 03



Lecture 07: a humanoid learns a K-pop challenge from phone video



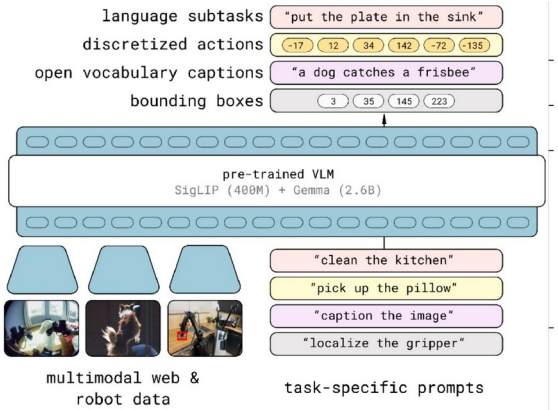
Lecture 03: robot actions need a coordinate system

Copying one dance is retargeting. Tonight: learning many tasks, with objects, from people's video.

Under the hood: what is inside a VLA?

Robot foundation models

Most commonly: a vision-language-action (VLA) model



- start with pretrained vision-language model (VLM)
- alternative design: start with generative video model
- training data mixture often includes:
  - robot demonstration data (imitation learning)
  - VLM tasks (question answering, captioning, detection)
  - human video data (motion prediction)
- often includes an diffusion-based "action expert"

5

VLA = VISION-LANGUAGE-ACTION

Start from a vision-language model (it already knows objects and words)

Add an "action expert": diffusion or flow matching (Lecture 03) outputs continuous joint commands

Train on robot demos + web tasks + human video

Output: a short chunk of actions, ~50 at a time

Adapted from Stanford CS 224R Deep RL, Lecture 17 "RL for Robot Foundation Models" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

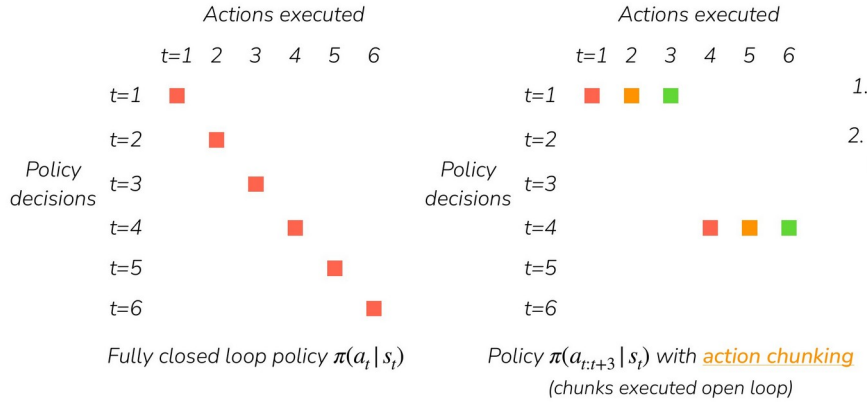
Under the hood: why predict a chunk of actions?

One more trick

So far: learning policy $\pi(a_t \mid s_t)$

—> *If timestep frequency is 50 Hz, then making new decisions every 20 ms.*

Alternative: predict a chunk of actions $a_{t:t+k}$ *based on state* s_t *and make new decisions less frequently (every* k *steps).*



Why?

1. It often works way better

2. Allows more computation time for policy inference

ACTION CHUNKING

A 50 Hz robot must decide every 20 ms

Predict the next ~1 second of actions at once

Smoother motion, fewer chances to hesitate

Same idea as generating a video clip, not a single frame

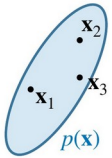
Adapted from Stanford CS 224R Deep RL, Lecture 2 "Imitation Learning" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

Under the hood: small mistakes compound

What can go wrong in imitation learning?

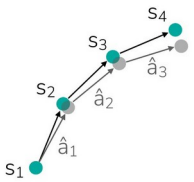
Compounding errors

Supervised learning



Inputs independent of predicted labels $\hat{\mathbf{y}}$

Supervised learning of behavior



Predicted actions affect next state.

Errors can lead to drift away from the data distribution!

Errors can then compound!

$$p_{expert}(\mathbf{s}) \neq p_{\pi}(\mathbf{s})$$

states visited by expert

states visited by learned policy π

"covariate shift"

23

COVARIATE SHIFT

The robot copies the expert, but never perfectly

One small error → a state the expert never showed

There it makes a bigger error → drift

Same shape as long-video drift (Lecture 07)

Adapted from Stanford CS 224R Deep RL, Lecture 2 "Imitation Learning" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

The data problem: can robots learn from videos of people?

Can robots directly use videos of people, animals?

Embodiment gap:
- difference in appearance
- difference in physical capabilities, degrees of freedom

Hard to directly imitate human & animal data, but can potentially guide exploration.

Peng, Kanazawa, Malik, Abbeel, Levine. SFV: Reinforcement Learning of Physical Skills from Videos. SIGGRAPH Asia 2018.

29

THE EMBODIMENT GAP

Robots have no internet of their own

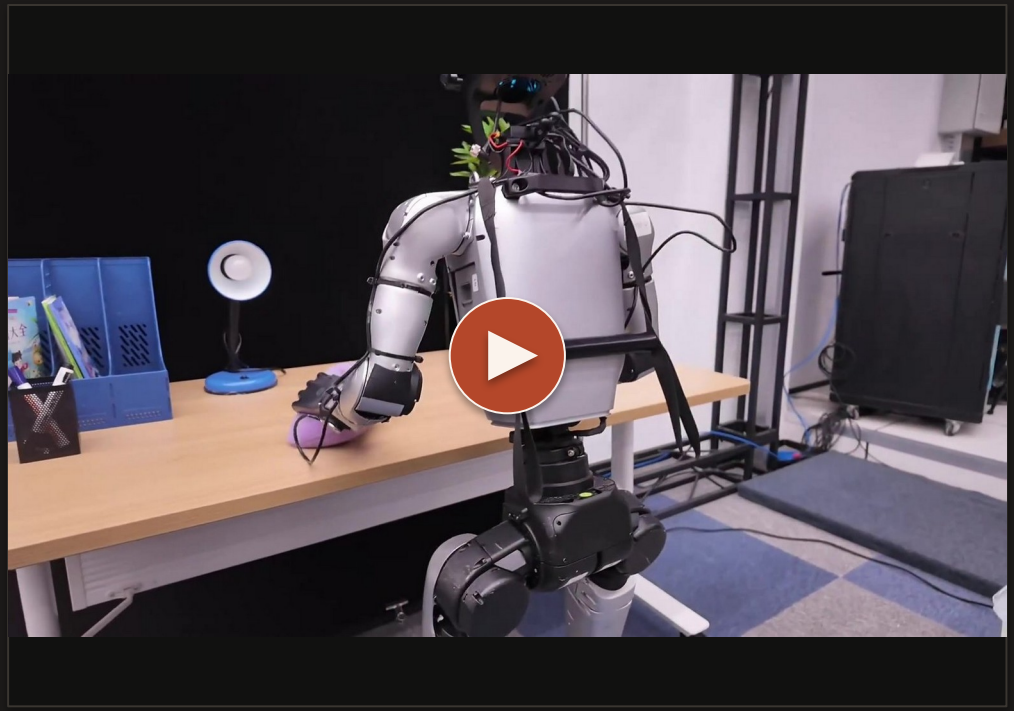
Teleoperation is precise but slow and costly

Human video is everywhere, but people look and move differently

Bridging this gap is tonight's main news

Adapted from Stanford CS 224R Deep RL, Lecture 2 "Imitation Learning" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

λ0 (26 Oct 2026): a humanoid that learned from people

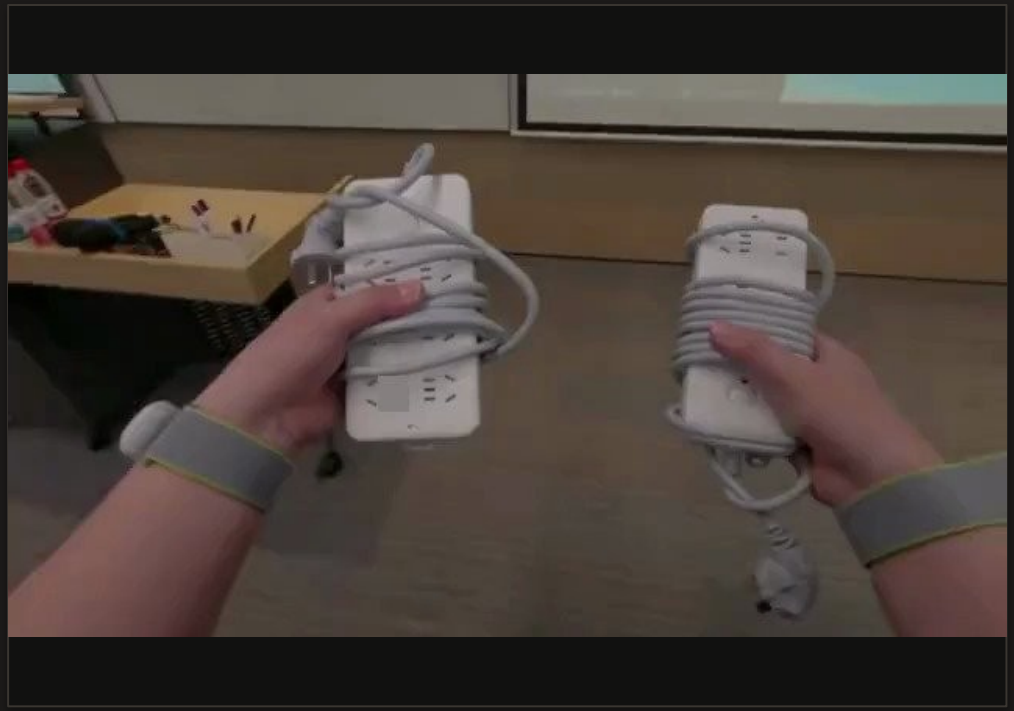


- From Alibaba's Amap and Token Hub, released yesterday
- Whole-body tasks: walk, bend, carry, use both hands
- Robot: Unitree G1 with five-fingered hands
- Trained mostly on HumanVerse: 500 hours of people doing everyday things

▶ λ0 project page

Alibaba Amap, λ0 project page (lambdahumanoid.github.io); TechNode, 26 October 2026.

HumanVerse: wear the robot’s eyes

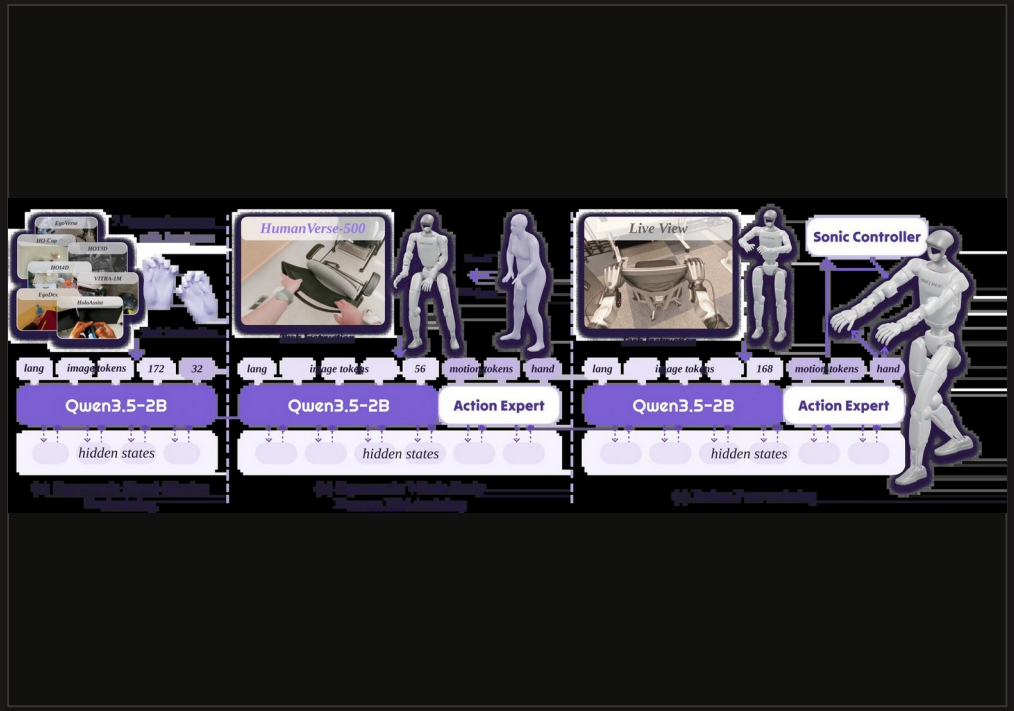


- A GoPro on the neck, at the robot’s camera height
- Body tracking: 24 joints; hands: 21 joints each
- Recorded in real places, not a lab
- About 3.4× faster to collect than teleoperation

▶ Watch the demo

λ0 project page: HumanVerse capture setup.

λ0: three stages, from human hands to a robot body



- I · Learn hands and objects from 7 first-person human datasets
- II · Translate HumanVerse into the robot's view and joints ("alignment")
- III · Finish on 600 real robot demos (150 per task)
- At run time: camera + instruction + body sense → whole-body motion

▶️ **Watch the demo**

λ0 project page: training recipe.

λ0, by the numbers

40.0%

success of the strongest baseline (π0.5) on 4 real tasks

62.5%

λ0: 25 of 40 real trials

22.5%

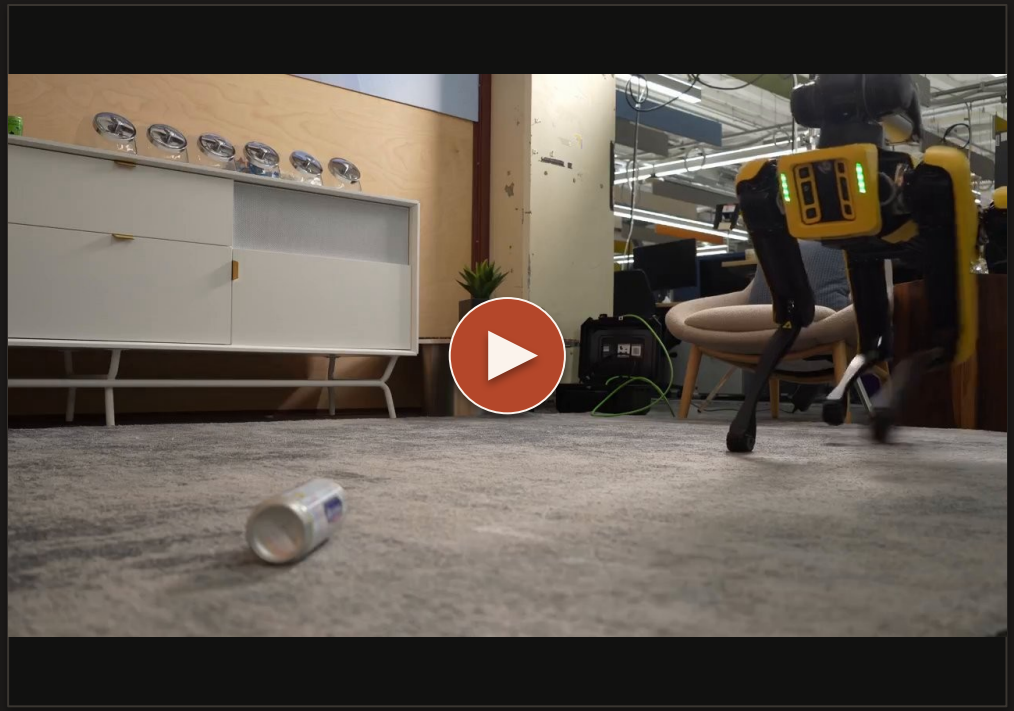
λ0 without stage II (no human-to-robot translation)

15.0%

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.

Gemini 4 Argon (13 Oct): watch a video, write a simulator

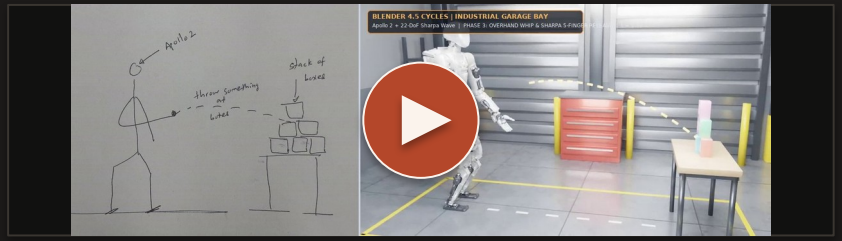


- A general Gemini model with strong physical reasoning
- Turns robot camera feeds, human video, even a sketch into a MuJoCo simulation
- Writes control code; practises in sim; improves real success (56% → 75% in one demo)
- Available in the Gemini API and AI Studio

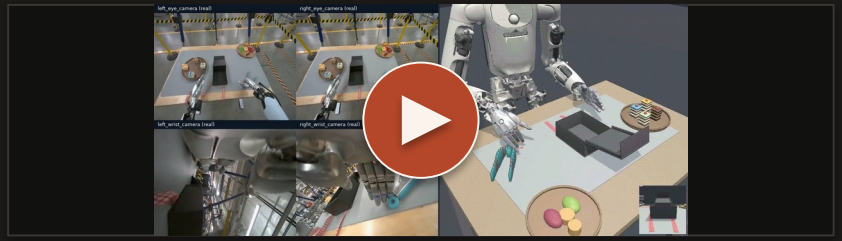
▶ **Argon robotics demos**

Google for Developers, “Gemini 4 Argon for robotics”, 13 October 2026.

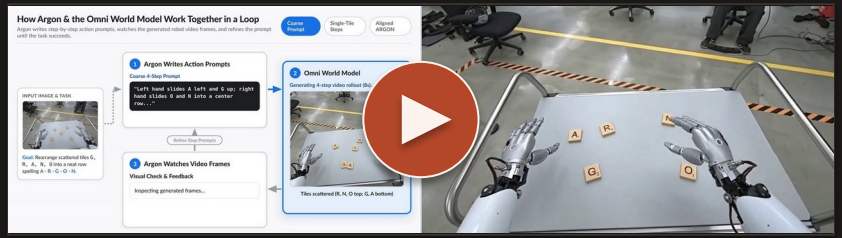
Argon demos: video → simulation → robot



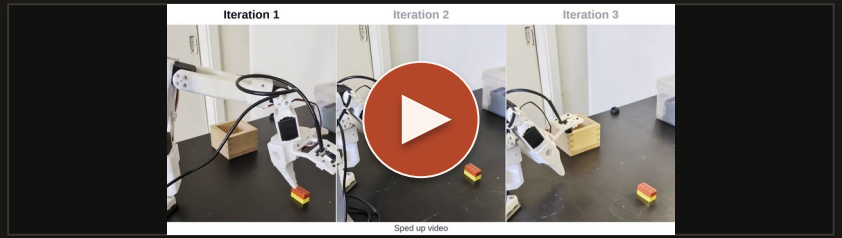
Sketch → simulation of a humanoid throwing a ball



Robot camera feeds → a calibrated simulated twin



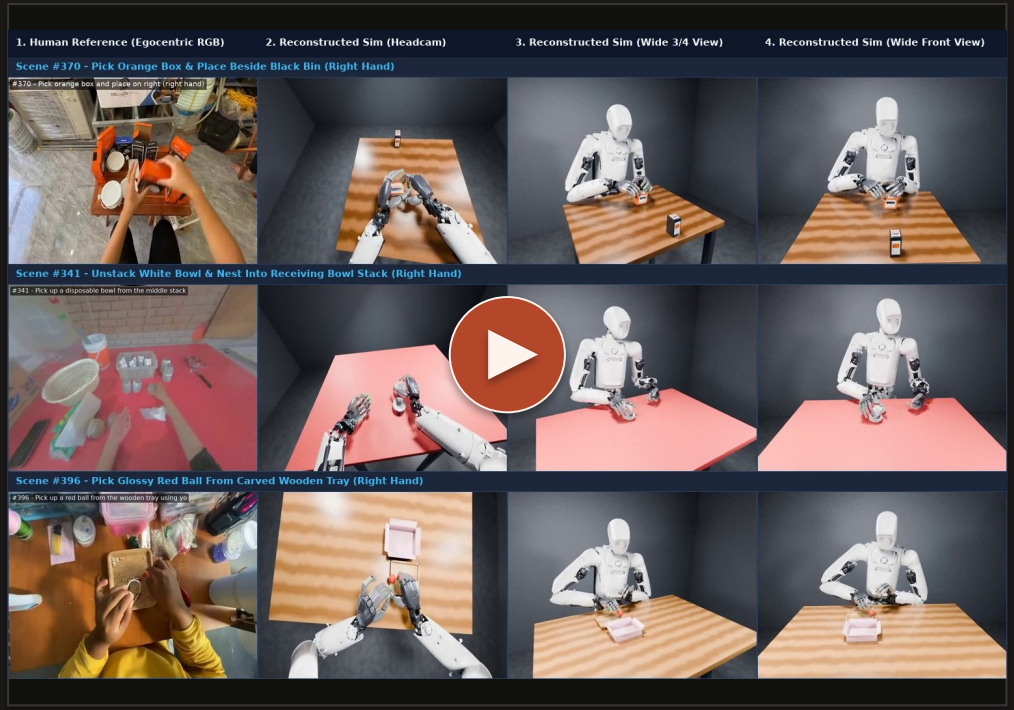
Steering a world model to spell A-R-G-O-N



SO-101 arm improves itself: 58.9 s → 22.9 s → 17.4 s

Google for Developers, “Gemini 4 Argon for robotics”, 13 October 2026 (demo videos).

Watch one person play cornhole, then do it

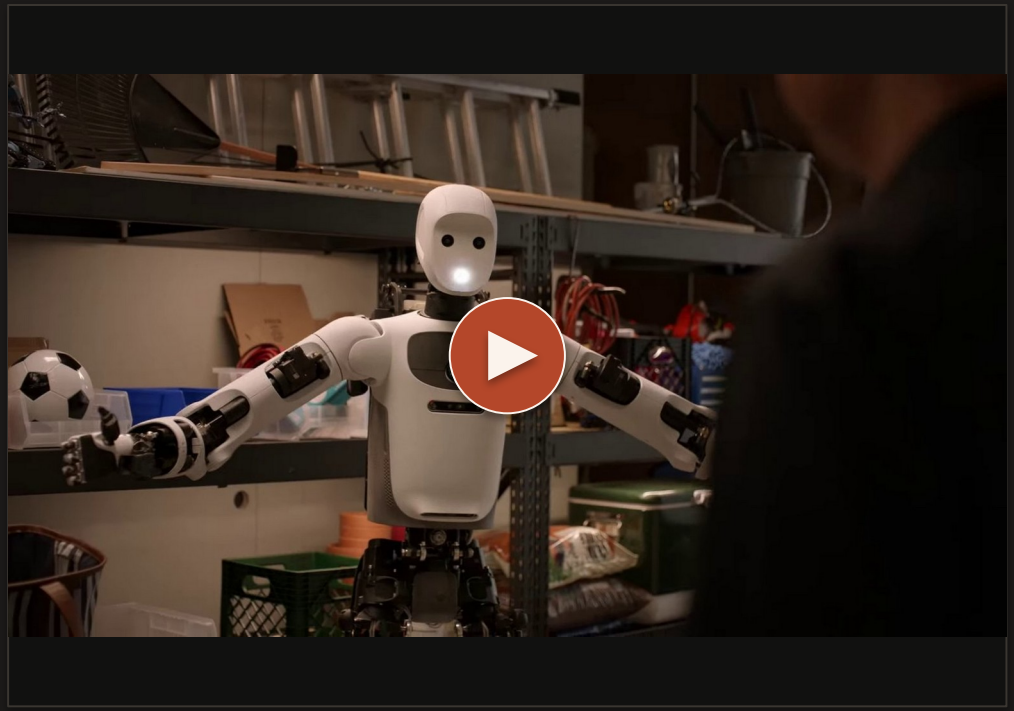


- Argon watches a raw video of a person playing cornhole
- Maps the space and works out whole-body joint motion
- Writes the simulation itself → trains → Apptronik Apollo
- Real trial: seven throws in a row on target

▶️ **Apollo cornhole (Oct 22)**

Robotics & Automation News, 22 October 2026, on Google DeepMind and Apptronik.

Gemini Robotics 2 (July): one model for the whole body

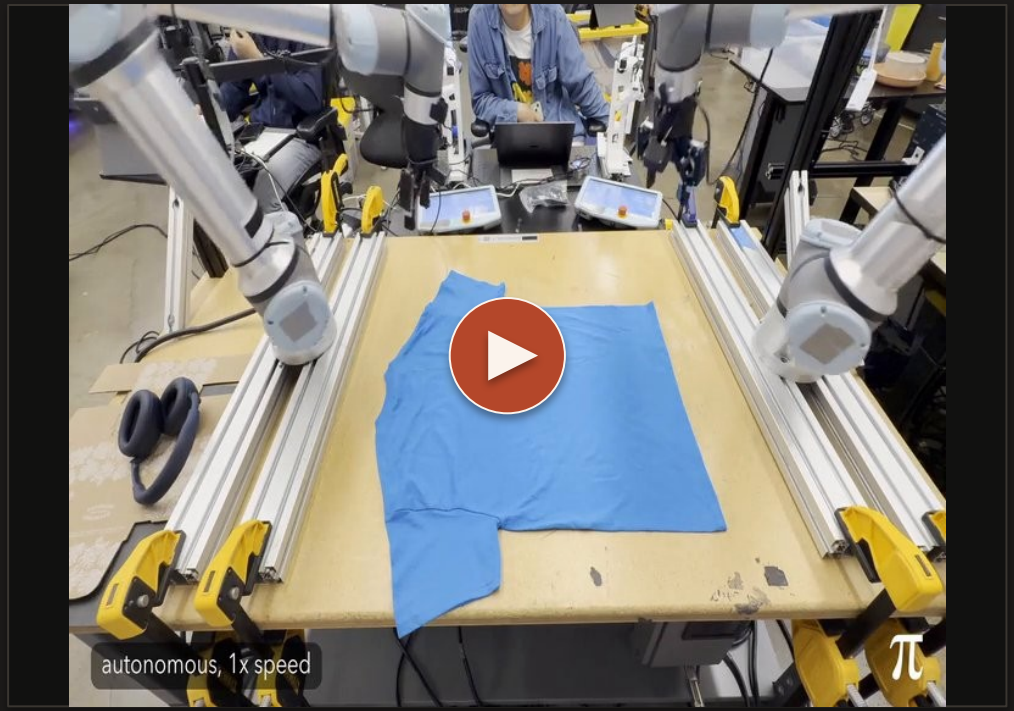


- VLA that walks and manipulates: “put the watering can in the green bin on the bottom shelf”
- ER 2 plans and checks progress; the VLA acts
- Adapts to a new robot body with fewer than ~200 examples
- Lecture 03, slide 86, in motion

▶ Gemini Robotics 2

Google DeepMind, Gemini Robotics 2, 30 July 2026; The Robot Report, 2 August 2026.

π0.7 and the Pi API: robot brains as a service



- π0.7 (April): steer with more than words: speed, quality, subgoal images
- Folded laundry on a robot with no laundry training data
- Pi API beta (13 Oct): upload demos → fine-tune → hosted inference
- Older π0 / π0.5 weights are open (openpi, Apache-2.0)

▶ Physical Intelligence

Physical Intelligence, π0.7 (16 April 2026) and Pi API beta (13 October 2026); openpi on GitHub.

Under the hood: why copying plateaus around 80%

What is the problem with VLAs?

Trained with *imitation learning* -> Performance often plateaus at around 80%

$\pi_{0.5}$ performance in unseen rooms

- similar to how LLMs become more performant with RL after SFT
- for robots to act *autonomously*, often need 99%+ reliability
  - -> natural use-case for RL-fine-tuning!
- pretrained VLA can serve as effective initialization for RL

Note: DAgger can also be helpful, often used in conjunction with RL

THE LAST 20%

Imitation-trained VLAs often plateau around 80% reliability

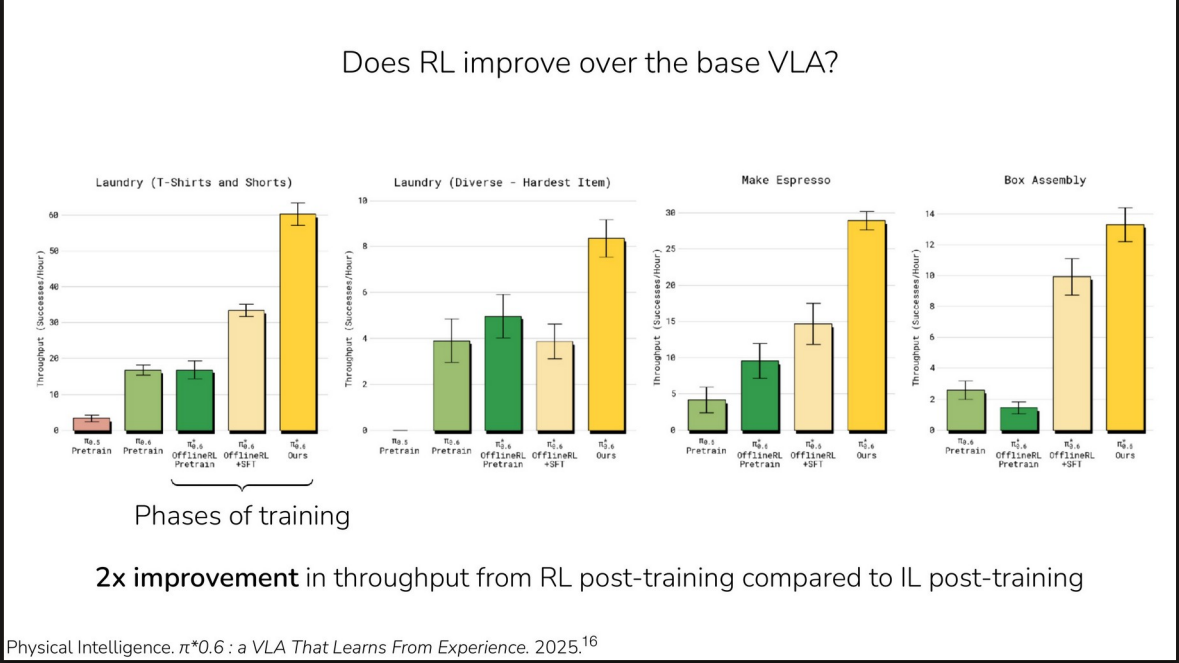
A home or factory needs 99%+

Copying can't be better than the demos

So: let the robot practise and learn from its own attempts (RL)

Adapted from Stanford CS 224R Deep RL, Lecture 17 "RL for Robot Foundation Models" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

Practice beats copying: RL roughly doubles throughput



Physical Intelligence. *π*0.6 : a VLA That Learns From Experience.* 2025.[16]

π*0.6 RESULT

Fit a value function: "is this going well?"

Fine-tune on what worked

About 2× throughput vs. imitation alone, on real tasks

Where world models help: practise in imagination, not on real dishes

Adapted from Stanford CS 224R Deep RL, Lecture 17 "RL for Robot Foundation Models" (Chelsea Finn, Spring 2026). cs224r.stanford.edu

Three ways to teach a robot, 2026

1

Steer it (teleoperation)

+ precise, the robot's own body
− slow, costly, needs a person per robot
π models: 10,000+ hours

2

Show it (human video)

+ cheap, everywhere, real homes
− different body, different eyes
λ0: 500 hours of HumanVerse

3

Imagine it (simulation, world models)

+ unlimited practice, rare events
− the sim-to-real gap (Lecture 07)
Argon, Cosmos, Atlas, Genie

Every serious system in 2026 mixes all three.

GAME · 4 MINUTES

Be the VLA

One student is the robot: eyes closed, moves only on instructions

A partner gives one-sentence commands: “move your right hand 20 cm left”

Goal: put a pen into a cup. Count the commands. Now try with eyes open

DISCUSSION · 3 MINUTES

If robots learn from first-person video of people's daily lives, whose bodies and homes become the training data?

Who should be paid or credited?

What should never be recorded?

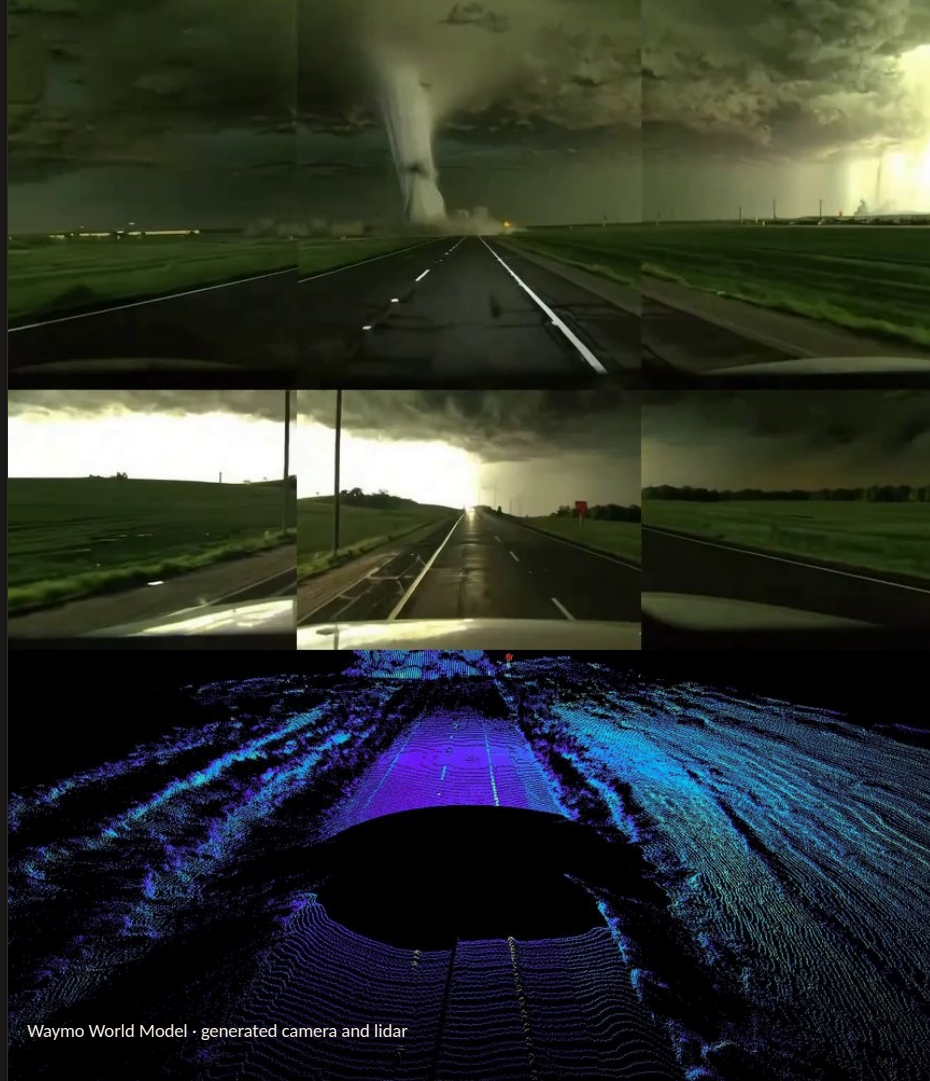
Would you wear the camera for a week? For how much?

STORY 4 OF 4

04

Does a world model need to draw pixels?

Generate the world, or predict it in a space you never see



Waymo World Model · generated camera and lidar

Two camps

A

Generate the world

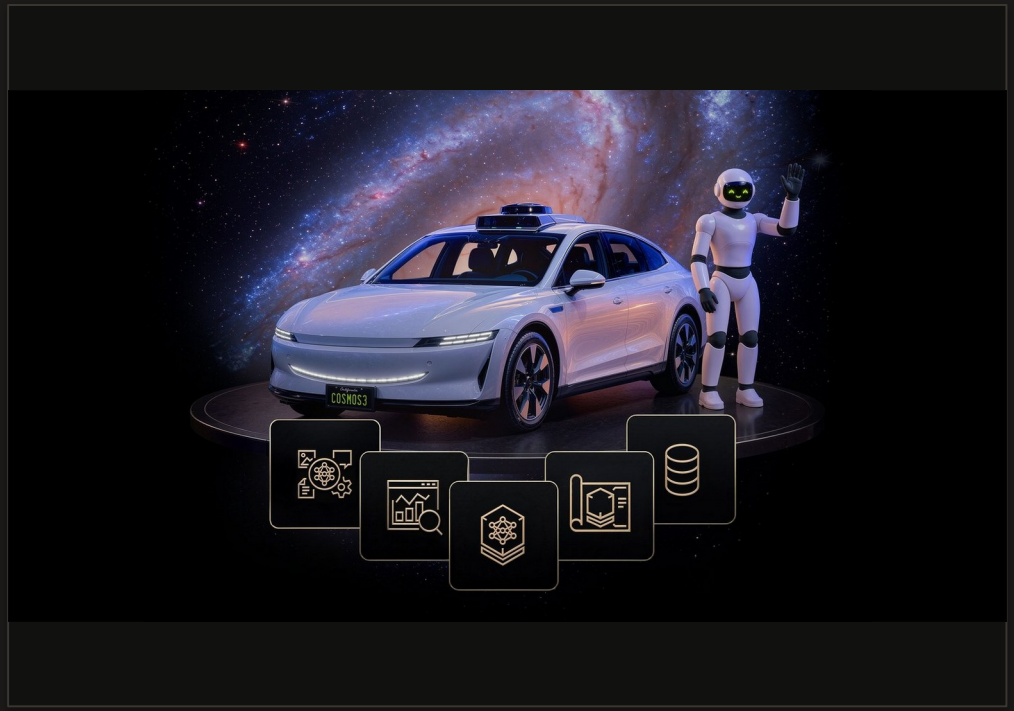
Predict the next frame in pixels (or splats). You can watch the prediction, train robots inside it, and sell it as video.
Genie 3 · Gemini Omni · Cosmos 3 · Atlas

B

Predict in latent space

Predict an abstract summary of the next moment, never the pixels. Ignore leaves and reflections; keep what matters for acting.
Yann LeCun · JEPA · V-JEPA 2 · AMI Labs

NVIDIA Cosmos 3: one open model to reason, simulate and act

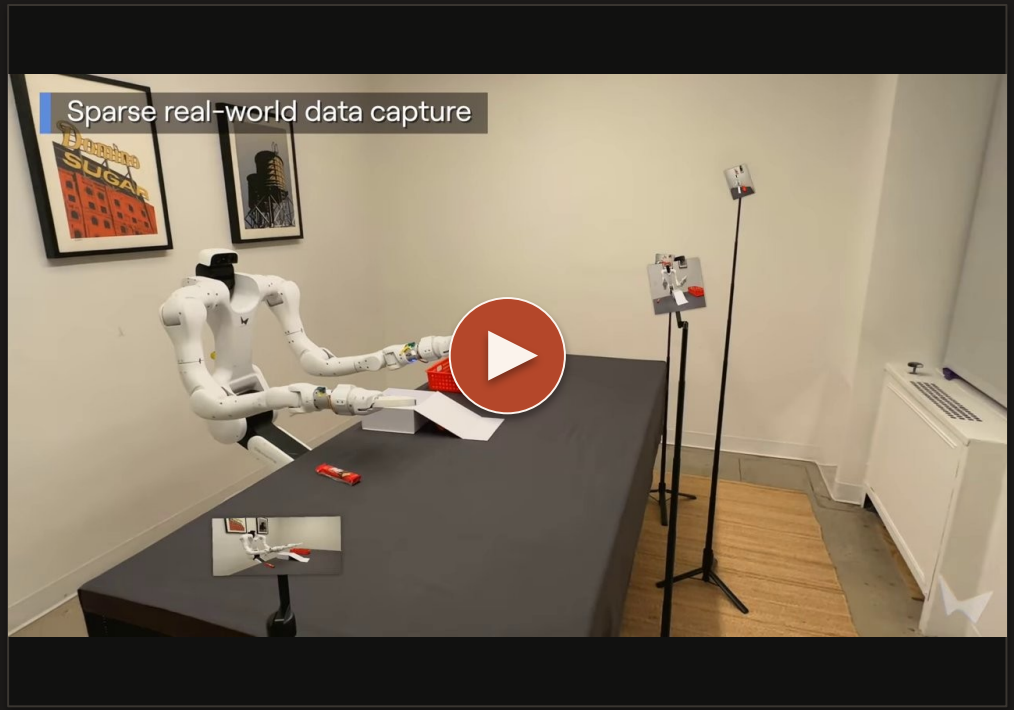


- Announced 31 May 2026 at GTC Taipei; fully open weights, data and training code
- Two towers: a Reasoner (VLM) + a Generator (video and actions)
- Super 32B for data centres; Nano 8B for a workstation
- Use: synthetic training data and testing robot policies

▶ NVIDIA Cosmos

NVIDIA, Cosmos 3 announcement, GTC Taipei, 31 May 2026; DataNorth summary, June 2026.

Why generate pixels? Because cameras are what robots see



- A robot's sensors see pixels and depth, so a simulator must produce them
- Atlas: 24 phone frames per room → RGB + depth for any robot path
- Waymo: rehearse crashes that can't be filmed
- Pixels are checkable: you can watch what the model believes

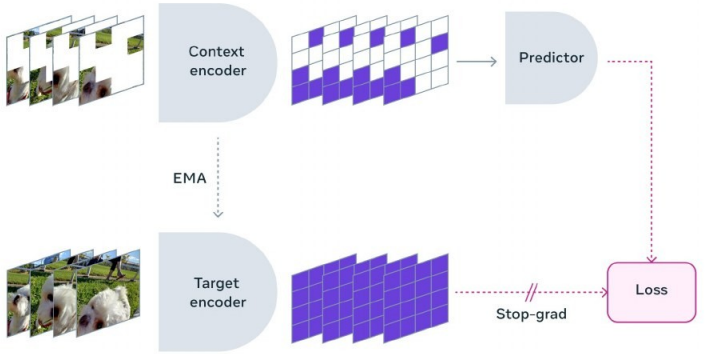
▶️ **Atlas robotics demo**

World Labs, "Atlas" (robotics section), September 2026; Waymo World Model.

Why not pixels? JEPA predicts the meaning, not the picture

V-JEPA 1-2

- A **non-generative** world model
- **Key idea:**
  - always work in **latent space**
  - minimize loss between:
    - the latent representation of true masked regions from raw video AND
    - the latent representation of the predicted masked regions from masked video
  - the loss itself is between the latents E(x) and E(y)
- **Goal:** to transfer latent representation to other tasks



$$\text{minimize}_{\theta,\phi,\Delta_y} \quad \|P_\phi(\Delta_y, E_\theta(x)) - \text{sg}(E_{\bar{\theta}}(y))\|_1,$$

Figure from https://arxiv.org/abs/2506.09985

Figure from https://ai.meta.com/blog/v-jepa-yann-lecun-ai-model-video-joint-embedding-predictive-architecture/

43

V-JEPA (META)

Mask part of a video; predict the representation of the missing part, not its pixels

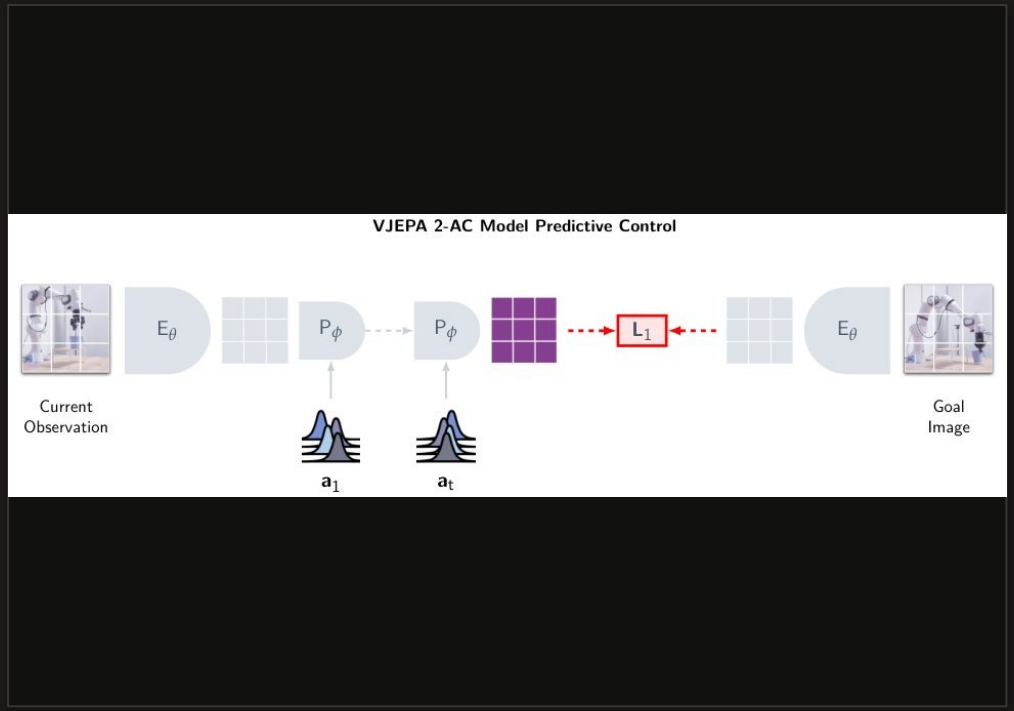
V-JEPA 2: pretrained on 1M+ hours of video

+ under 62 hours of robot video → an action-conditioned model that can plan

LeCun's AMI Labs: $1.03B seed (March 2026) to build this

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423 Facts: Meta V-JEPA 2 (2025); TorontoStarts on AMI Labs, Sept 2026.

V-JEPA 2-AC: plan a robot move without drawing the future



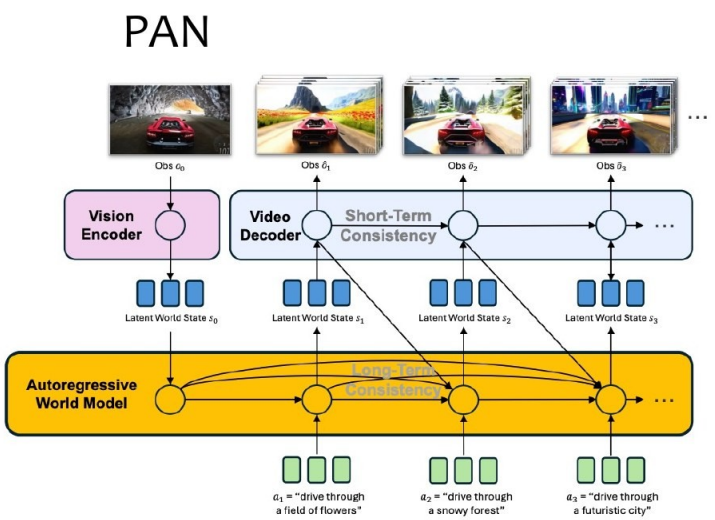
- Give a goal image: “the cup should end up here”
- Imagine many action sequences, in embedding space only
- Pick the one whose predicted embedding is closest to the goal
- Execute a step, look again, re-plan

▶ V-JEPA 2 paper

Assran et al., “V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning”, Meta, 2025 (arXiv 2506.09985).

Under the hood: the two losses, side by side

- Goal: interactive world model for long-horizon, conditioned simulation
- Model: three parts
  - *Vision Encoder (h):* maps observations into structured latent states
  - *Autoregressive World Model (f):* predicts next latent state conditioned on actions & history (**long horizon**)
  - *Video Diffusion Decoder (g):* reconstructs high-quality visual observations from latent states (**short horizon**)



$$p_{\text{PAN}}(o_{t+1} \mid o_t, a_t) = \sum_{\hat{s}_t, \hat{s}_{t+1}} \underbrace{p_h(\hat{s}_t \mid o_t)}_{\text{encoder}} \underbrace{p_f(\hat{s}_{t+1} \mid \hat{s}_t, a_t)}_{\text{world model}} \underbrace{p_g(o_{t+1} \mid \hat{s}_{t+1})}_{\text{decoder}}$$

Figure from https://arxiv.org/pdf/2511.09057

45

READ THE TWO LINES

Generative: make the predicted next observation look like the real one (pixels)

JEPA: make the predicted embedding match the real embedding

Same structure: state + action → next

The fight is about what to compare: pictures or meanings

Adapted from CMU 10-423/623/723 Generative AI, Lecture 26 "Interactive World Models" (Matt Gormley & Aran Nayebi). cs.cmu.edu/~mgormley/courses/10423

How would you know a world model really “understands”?

1

Round trip

Walk away and come back: is it the same? (Lecture 03)

2

Counterfactual

Push the cup instead of lifting it: does the result change correctly?

3

Long horizon

Still consistent after 10 minutes? An hour?

4

Real action

Does a robot that planned inside it succeed in the real room?

Only the last test is the same for both camps. That is why robots have become the referee.

DEBATE · 10 MINUTES

“A video generator is a world model.”

Left side of the room argues for; right side against

Two speakers per side, one minute each

Then the whole room votes by show of hands

Arguments to start from

+

For

- Genie 3 already takes actions and keeps a minute of memory
- Waymo trains real cars with it
- Argon steers Omni to move Scrabble tiles: it predicts consequences
- To draw the future correctly, you must model the world

−

Against

- Most video models take no actions: they predict, they don’t respond
- They forget, drift and break physics (Lecture 07)
- Pretty ≠ correct: a beautiful wrong future
- JEPA: the world is not its pixels

Four stories, one line each

01

Playable worlds

Add an action to a video model and it becomes a place; the hard part is memory

02

Splats

Keep the geometry and the world becomes a file you can open, edit and sell

03

Bodies

Robots now learn mostly from people and from imagined worlds, but still fail 1 time in 3

04

Pixels?

Whether a world model must draw the world is still open, and robots will decide

Tonight → your assignments (both due Nov 30)

A3

Assignment 3: your game

Sketch a level in a world model or Marble; or scan a real place as a splat and load it in three.js. Use the “directing a world” checklist

A2

Assignment 2

A splat or a world-model clip can be a set; disclose which tools made which parts

→

Final project ideas

An explorable memory · a world you can only enter once · a robot that learns your gesture

Three readings for this week

1

Atlas

World Labs (Sept 2026). One model for images, video and 3D; read the “spatial context” and robotics sections

Technique · worldlabs.ai/blog/atlas

2

λ0

Alibaba Amap (Oct 2026). Read the HumanVerse and three-stage training sections, and look at the ablation

Robotics · lambdahumanoid.github.io

3

Genie 3 and the AGI question

TechCrunch (Aug 2025). How DeepMind frames world models as a step towards general AI

Culture · techcrunch.com

Quiz 09 · five short questions

1

Q1 · Technique

What one input turns a video model into a world model?

2

Q2 · Technique

Why do Gaussian splats render so much faster than NeRF?

3

Q3 · Art

Choose a place you would scan as a splat. What would the copy keep and lose?

4

Q4 · Art

Whose daily life should (not) become robot training data?

5

Q5 · Reading

One claim from a reading you would want to test, and how

Every demo from tonight, in one place

- ▶️ Genie 3 (Google DeepMind)
- ▶️ Project Genie
- ▶️ Waymo World Model
- ▶️ Gemini Omni
- ▶️ GameNGen: DOOM in a diffusion model
- ▶️ World Models (Ha & Schmidhuber, 2018)
- ▶️ Marble 1.1 (World Labs)
- ▶️ Atlas (World Labs)
- ▶️ AMD to acquire World Labs
- ▶️ HY-World 2.0 (Tencent)
- ▶️ 3D Gaussian Splatting (Inria)
- ▶️ Splats in Sept 2026 (Radiance Fields)
- ▶️ λ0 (Alibaba Amap)
- ▶️ Gemini 4 Argon for robotics
- ▶️ Apollo plays cornhole
- ▶️ Gemini Robotics 2
- ▶️ Physical Intelligence π0.7 / Pi API
- ▶️ NVIDIA Cosmos 3
- ▶️ Meta V-JEPA 2
- ▶️ CMU 10-423 Lecture 26 (World Models)
- ▶️ Stanford CS231A Lecture 17 (Splatting)
- ▶️ Stanford CS 224R (VLAs, imitation)

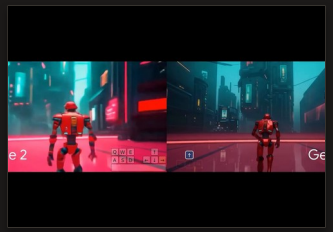
NEXT WEEK

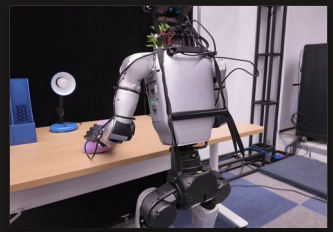
Critique, evaluation, provenance, and responsible release

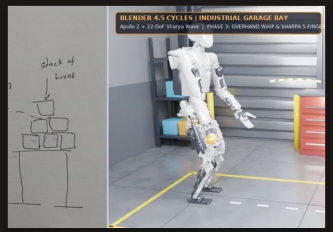
Now that we can make worlds, voices and bodies: how do we judge them, label them, and decide what to release?

AMCC 5160 · LECTURE 09 · 27 OCTOBER 2026

Worlds, Splats, *and Bodies*







Thank you · Questions?

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

APPENDIX

DIY

Do it yourself

Tools, a one-evening workflow, and a checklist

3D Gaussian Splatting · Inria

Toolkit: what to use for each job

Job	Tool	Runs on	Open?
World from text or one image	Marble 1.1 (World Labs) · HY-World 2.0 (Tencent)	Web, credits · 1 large GPU	No · Yes
Try a playable world	Project Genie (Google, US Ultra) · Gemini Omni Flash	Web / app	No
Scan a real place	Postshot · Brush · nerfstudio gsplat · Polycam, Scaniverse	Laptop GPU or phone	Mixed
Edit and clean splats	SuperSplat 3.0 (PlayCanvas)	Browser	Yes
Use splats in a scene	Blender 5.3 (alpha) · three.js r186 · Spark	Laptop / browser	Yes
Robot policy	openpi (π0, π0.5) · LeRobot + SO-101 arm	>8 GB GPU for inference	Yes
Simulate	MuJoCo · NVIDIA Isaac Sim · Cosmos 3 Nano (8B)	Laptop · workstation GPU	Yes

Check each licence before commercial use. Sources: vendor pages and release notes, Sept–Oct 2026.

Workflow: a walk-in level for your game, in one evening

1 · Pick a place (real or prompted) → 2 · Scan 60–150 photos, or generate in Marble / HY-World → 3 · Clean in SuperSplat; set scale and floor

4 · Load in three.js / Spark or Blender → 5 · Add a collider and one interaction → 6 · Playtest: walk away, come back, break it

Budget: ~30 min capture, ~40 min training, ~30 min cleaning, ~1 h in the engine. Keep the original photos and the export settings (Lecture 07: keep your sources).

Review checklist for a generated world

1

Consistency

Round trip passes? Objects keep colour, size, place?

2

Scale and physics

Doors at human height? Things fall the right way?

3

Invented parts

Which parts did the model make up? Are they labelled?

4

Rights

Did you have permission to scan? Whose photos?

5

Export

Format, splat count, file size, licence of the tool

6

Disclosure

Credits: which model made which part (Week 10)